FINAL REPORT

ESTCP Pilot Program
Classification Approaches in Munitions Response

November 2008

Herb Nelson,
ESTCP

Katherine Kaye,
ESTCP Support Office, HydroGeoLogic, Inc.

Anne Andrews,
ESTCP

Environmental Security Technology
Certification Program

1 REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0811

2 PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
17-11-2008	**Final Report**	2007- 2008

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ESTCP Pilot Program- Classification Approaches in Munitions Response	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Herbert Nelson, ESTCP **Katherine Kaye, ESTCP Support Office (HydroGeoLogic, Inc.)** **Anne Andrews, ESTCP**	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES	8. PERFORMING ORGANIZATION REPORT NUMBER
ESTCP 901 North Stuart Street, Suite 303, Arlington, VA 22203	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITORS ACRONYM(S)
Environmental Security Technology Certification Program 901 North Stuart Street, Suite 303 Arlington, VA 22203	ESTCP
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The first demonstration of the ESTCP classification pilot program was conducted on the former Camp Sibert, AL. This site was used for 4.2-inch mortar and has generally benign topography and vegetation, allowing the collection of high-quality geophysics data, and benign to moderate geology. Data were collected with four systems: a magnetometer, an EM61-MK2 cart, and EM61-MK2 array, and a next-generation electromagnetic sensor (BUD). Researchers applied classification algorithms to these data to make a determination about whether each item detected was likely to arise from a munitions item or clutter object. All the detected objects were carefully excavated to allow for algorithm training and blind testing of the classification approaches.

The pilot program demonstrated successful classification on this simple site. With carefully collected survey data and transitioning physics-based analysis techniques, well over half the detected clutter items were routinely eliminated with high confidence, while retaining all the munitions. In all cases, the classification processing correctly identified all or nearly all the munitions and a significant fraction of the clutter was successfully identified as such with high confidence. Classification processing, applied to data from the commercial instruments, eliminated 45%–70% of the clutter in these examples. When advanced emerging EM sensors were used, nearly perfect results were achieved.

15. SUBJECT TERMS

Environmental Security Technology Certification Program (ESTCP)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	3 18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U			

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

ACKNOWLEDGMENTS

The ESTCP Classification Pilot Program would not have been possible without the assistance of numerous individuals. We would like to acknowledge Shelley Cazares, Michael Tuley, and Michael May from the Institute for Defense Analyses who were instrumental in the overall program design and analysis of the demonstrator results. We would also like to thank the ESTCP Classification Study Advisory Group listed below. They were involved in site selection, program design, data review, and the development of conclusions and methods as a whole.

James Austreng, California Department of Toxic Substances Control
Steve Cobb, Alabama Department of Environmental Management
Harry Craig, U.S. EPA Region 10, Oregon Operations Office
Jon Haliscak, AFCEE Technical Directorate
David Patterson, AFCEE Technical Directorate
Andrew Schwartz, U.S Army Corps of Engineers, Huntsville
Robert Selfridge, U.S. Army Corps of Engineers, Huntsville
Tracy Strickland, Alabama Department of Environmental Management
Jeff Swanson, Colorado Department of Public Health and Environment
Ken Vogler, Colorado Department of Public Health and Environment
Roger Young, U.S. Army Corps of Engineers, Huntsville

We would like to credit the technology demonstrators.

- SAIC, Inc., Signal Innovations Group, and Sky Research Inc. for classification data analysis;
- Lawrence Berkeley National Laboratory for data collection and supporting analysis for the Berkeley Unexploded Ordnance Discriminator (BUD);
- Nova Research, Inc. for the EM-array, magnetometer array and GEMTADS data collections;
- Parsons for the EM61-MK2 cart data collection and mag & flag demonstration, in addition to general onsite support;
- NAEVA Geophysics Inc. for the cued GEM3 data collection; and
- Sky Research Inc. for EM63 data collection and supporting data analysis.

For program support, we acknowledge

- Gregory Nivens for the valuable onsite support provided by him and his team from Parsons;
- Sheri Anderson-Hudgins and Bob Selfridge from the USACE for their assistance in the Former Camp Sibert demonstrations;
- Nagi Khadr from SAIC, Inc. who served as a representative for the ESTCP Program Office, for data analysis; and
- Clif Youmans, Montana Army National Guard, for supplying the recovered inert rounds used for the seeding.

Finally, ESTCP thanks the Kell and the Freeman families, owners of the land involved in this demonstration, for their patience and cooperation.

TABLE OF CONTENTS

Executive Summary 1
1 Introduction 3
1.1 Background 3
1.2 Classification Concept 3
1.3 ESTCP Pilot Program 5
1.4 About This Report 6
2 Former Camp Sibert 7
2.1 Test Site Objectives 7
2.2 Site History and Characteristics 7
2.3 Study Site Overview 8
2.4 Demonstration Preparation 8
3 Program Design 11
3.1 Overall Approach 11
3.2 Evaluating Success 11
3.3 Data Collection 12
3.4 Classification Approaches 15
3.5 Ground Truth 17
3.6 Classification Product 18
3.7 Scoring Methods 19
4 Detection 21
4.1 Anomaly-Selection Threshold 21
4.2 Master Anomaly List 23
4.3 Mag and Flag Detections 24
5 Classification Results 25
5.1 Feature Extraction 25
5.2 Algorithm Training 26
5.3 Blind Test Results 28
5.4 Determination of Threshold 33
6 Data Requirements 35
6.1 Density 35
6.2 Geolocation Precision 36
6.3 Signal-to-Noise Ratio 37
6.4 Data Acquisition Observations at Sibert 38
7 Cost Considerations 41
7.1 Cost Model 41
7.2 Cost Drivers 44
8 Program Conclusions 47
8.1 Overall 47
8.2 Ability to Correctly Determine Parameters 47
8.3 Understanding Failures 47
8.4 Site–Specific Factors Affecting Performance 48
9 Classification Implementation 51
9.1 Practical Model for the Classification Process 51
9.2 Applications of Classification to the Munitions Response Process 52
9.3 Factors Affecting Acceptance of Classification 53

10 Frequently Asked Questions About Classification 55
References 61

Executive Summary

The detection and remediation of munitions is one of the Department of Defense's (DoD) most pressing environmental problems. The Military Munitions Response Program (MMRP) is charged with characterizing and, where necessary, remediating munitions-contaminated sites. When a site is cleaned up, it is typically mapped with a geophysical system, based on either a magnetometer or electromagnetic (EM) induction sensor, and the locations of all detectable signals are excavated. Many of these detections do not correspond to munitions, but rather to other harmless metallic objects or geology. Application of technology to separate the munitions from other objects, known as classification, offers the potential of significant cost savings for munitions response.

Field experience indicates that often in excess of 90% of objects excavated during the course of a munitions response are found to be nonhazardous items. Current technology, as it is commonly implemented, does not provide a physics-based, quantitative, validated means to discriminate between hazardous munitions and nonhazardous items. With no information to suggest the origin of the signals, all anomalies currently are carefully excavated by certified unexploded ordnance (UXO) technicians using a process that often requires expensive safety measures, such as barriers or exclusion zones. As a result, most of the costs to remediate a munitions-contaminated site are currently spent on excavating targets that pose no threat. If these items could be determined with high confidence to be nonhazardous, some of these expensive measures could be eliminated or the items could be left unexcavated entirely.

Classification is a process used to make a decision about the likely origin of a signal. In the case of munitions response, high-quality geophysical data can be interpreted with physics-based models to estimate parameters that may be useful for classification. The parameters in these models are related to the physical attributes of the object that resulted in the signal, such as its physical size and aspect ratio. The geophysical data can be analyzed to estimate the values of these parameters, which may then be used to estimate the likelihood that the signal arose from an item of interest, that is, a munition.

The first demonstration of the ESTCP pilot program was conducted on the former Camp Sibert, AL. This site was used for 4.2-inch mortar training in World War II. It has generally benign topography and vegetation, allowing the collection of high-quality geophysics data, and benign to moderate geology. Data were collected with several commercial and emerging magnetic and electromagnetic sensors, and researchers applied classification algorithms to these data to make a determination about whether each item detected was likely to arise from a munitions item or clutter object. All the detected objects were carefully excavated to provide ground truth information to allow for algorithm training and blind testing of the classification approaches. The main objective was for the classifiers to correctly identify with high confidence the detected objects that were not hazardous. As such, the main failure was considered to be any error where a munitions object was declared with high confidence to be nonhazardous.

The pilot program demonstrated successful classification on this simple site. With carefully collected survey data from either magnetometers or EM sensors and transitioning physics-based analysis techniques, well over half the detected clutter items were routinely eliminated with high confidence, while retaining all the munitions. Table ES-1 shows example results from the Camp Sibert demonstration for classification algorithms applied to data collected with four systems: a

magnetometer, an EM61-MK2 CART, and EM61-MK2 array, and a next-generation EM sensor (BUD). In all cases, the classification processing correctly identified all or nearly all the munitions and a significant fraction of the clutter was successfully identified as such with high confidence. Classification processing, applied to data from the commercial instruments, eliminated 45%–70% of the clutter in these examples. When advanced emerging EM sensors were used, nearly perfect results were achieved.

Table ES-1. Example classification results from the Camp Sibert demonstration

Sensor/Performer	# Munitions	% Munitions Correctly Identified	# Non-munitions Detected	% Non-munitions Correctly Identified
Mag Array/Sky Research	118	100	615	44
EM61-MK2 Cart/Parsons	145	99	488	44
EM61-MK2 Array/SAIC, Inc.	119	99	615	72
BUD/Lawrence Berkeley National Laboratory	56	100	209	97

Realizing the potential advantages of classification requires formulating a model for its application that will be accepted by all stakeholders. The study described in this report relied on a retrospective analysis of demonstrator performance. Classification on a production site would need to proceed in a prospective rather than retrospective model. Truth information that would allow one to determine whether each individual item was correctly classified as munitions or clutter will not be available if the clutter is not dug up. The program advisory group considered how a site team might proceed with a dig program, specifically, how to decide when to stop digging when classification is used.

The MMRP is severely constrained by available resources. Remediation of the entire inventory using current practices is cost prohibitive, within current and anticipated funding levels. With current planning, estimated completion dates for munitions response on many sites are decades out. If the savings potential of applying classification technologies were realized, the limited resources of the MMRP could be used to accelerate the cleanup of munitions response sites that are currently forecast to be untouched for decades.

1 INTRODUCTION

1.1 BACKGROUND

Munitions response is a high-priority problem for the Department of Defense (DoD). Approximately 3000 sites, comprising tens of millions of acres, are suspected of contamination with military munitions, include unexploded ordnance (UXO) and discarded military munitions. (Ref. 1) The bulk of these are formerly used defense sites (FUDS), which are no longer under DoD control, and are used for a variety of purposes, including residential development, recreation, grazing, and parkland, often without restriction.

The Military Munitions Response Program (MMRP) is charged with characterizing and, where necessary, remediating munitions-contaminated sites. When a site is cleaned up, it is typically mapped with a geophysical system, based on either a magnetometer or electromagnetic (EM) induction sensor, and the locations of all detectable signals are excavated. Many of these detections do not correspond to munitions, but rather to other harmless metallic objects or geology, termed clutter: field experience indicates that often in excess of 90% of objects excavated during the course of a munitions response are found to be nonhazardous items. Current technology, as it is commonly implemented, does not provide a physics-based, quantitative, validated means to discriminate between hazardous munitions and nonhazardous items.

With no information to suggest the origin of the signals, all anomalies are currently treated as though they are intact munitions when they are dug. They are carefully excavated by certified UXO technicians using a process that often requires expensive safety measures, such as barriers or exclusion zones. As a result, most of the costs to remediate a munitions-contaminated site are currently spent on excavating targets that pose no threat. If these items could be determined with high confidence to be nonhazardous, some of these expensive measures could be eliminated or the items could be left unexcavated entirely.

The MMRP is severely constrained by available resources. Remediation of the entire inventory using current practices is cost prohibitive, within current and anticipated funding levels. With current planning, estimated completion dates for munitions response on many sites are decades out. The Defense Science Board (DSB) observed in its 2003 report that significant cost savings could be realized if successful classification between munitions and other sources of anomalies could be implemented. (Ref. 2) If these savings were realized, the limited resources of the MMRP could be used to accelerate the clean up of munitions response sites that are currently forecast to be untouched for decades.

1.2 CLASSIFICATION CONCEPT

Classification is a process used to make a decision about the likely origin of a signal. In the case of munitions response, high-quality geophysical data can be interpreted with physics-based models to estimate parameters that may be useful for classification. The parameters in these models are related to the physical attributes of the object that resulted in the signal, such as its physical size and aspect ratio. The values of these parameters may then be used to estimate the likelihood that the signal arose from an item of interest, that is, a munition.

Magnetometer data are typically fit using a simple model of a single dipole moment, which is related to the physical size of the object. EM data are fit to a more complex three-axis polarizability model that can yield a larger set of parameters that more completely describe the source of the signal. Figure 1-1 illustrates schematically the EM measurement. The three-axis response model is indicated by the heavy arrows in the right panel of the figure, which shows responses along the three perpendicular principal axes of the munition. The EM parameters relate to the physical size of the object, its aspect ratio, the wall thickness, and the material properties.

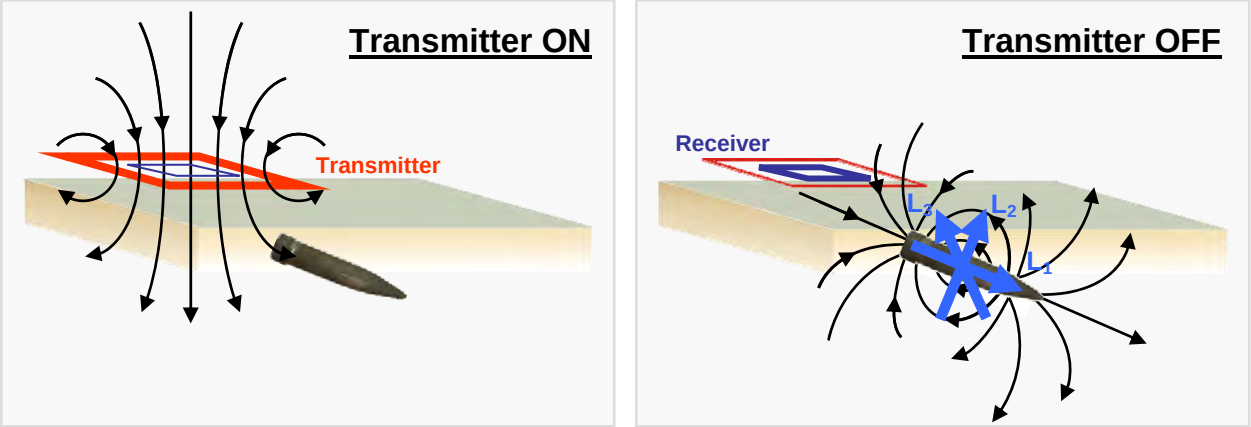

Figure 1-1. Schematic of EM measurement and three-axis electromagnetic response

Munitions are typically long, narrow cylindrical shapes that are made of heavy-walled steel. Common clutter objects can derive from military uses and include exploded parts of targets, such as vehicles, as well as munitions fragments, fins, base plates, nose cones and other munitions parts. Other common clutter objects are man-made nonmilitary items. While the types of objects that can possibly be encountered are nearly limitless, common items include barbed wire, horseshoes, nails, hand tools, and rebar. These objects and geology give rise to signals that will differ from munitions in the parameter values that are estimated from geophysics data.

Once the parameters are estimated, a means must be found to sort the signals to identify items of interest, in this case munitions, from the clutter. This is termed classification. In a simple situation, one can imagine sorting items based on a single parameter, such as object size. A rule could be made that all objects with an estimated size larger than some value will be treated as potentially munitions items of interest, such as large bombs, and those smaller could not possibly correspond to intact munitions.

In reality, few classification problems can be handled successfully based on a single parameter. Because the parameter-estimation process is imperfect and the physical sizes of the objects of interest may overlap with the sizes of the clutter objects, it is rare to get perfect separation based on one parameter. For complex problems, sophisticated statistical classifiers can combine the information from multiple parameters to make a quantitative estimate of the likelihood that a signal corresponds to an object of interest.

1.3 ESTCP PILOT PROGRAM

The Environmental Security Technology Certification Program (ESTCP) is charged with promoting innovative, cost-effective environmental technologies by demonstrating and validating those technologies. In response to the DSB Task Force report (Ref. 2) and Congressional interest, ESTCP initiated a Classification Pilot Program to validate the application of a number of recently developed technologies in a comprehensive approach to munitions response.

Some form of classification is used on all munitions response projects, most often implicitly. In the case of traditional "mag and flag," the operator adjusts the sensitivity audio control and makes a decision as to whether each signal is significant. Since no data are recorded, these decisions can never be reviewed. In the case of digital geophysical mapping, a threshold is selected for determining targets of interest, and often a geophysicist uses professional judgment to decide based on a visual inspection of shape and amplitude whether anomalies are likely to arise from geology or compact metallic objects. In both cases, the sources of signals deemed insignificant are not further investigated and remain in the ground.

Significant progress has been made in explicit classification technology. To date, emerging technologies have primarily been tested at constructed test sites, with only limited application at live sites. The routine implementation of classification technologies will require demonstrations at real munitions response sites under real-world conditions. Any attempt to declare detected anomalies to be harmless will require demonstration to regulators, safety personnel, and project managers of not only individual technologies, but an entire decision-making process.

The goal of the pilot program is to demonstrate that classification decisions can be made explicitly, based on principled physics-based analysis that is transparent and reproducible. As such, the objectives of the pilot program were to:

- Test and validate detection and classification capabilities of currently available and emerging technologies on a real site under operational conditions, and
- Investigate how classification technologies can be implemented in cleanup operations in cooperation with regulators and program managers.

To address the second of those objectives, a Program Advisory Group composed of representatives of the Services and State and National regulators was established at the beginning of the program. This Advisory Group was involved in site selection, program design, data review, and the development of conclusions. The Advisory Group has been heavily involved in drafting this report.

The Former Camp Sibert in Alabama was selected as the first pilot site with success in mind. This site presented a single munitions type (the 4.2-inch mortar) and benign conditions where high-quality data could be collected. The motivation of this selection was to demonstrate a process under conditions where the technologies were expected to perform well, so that the advisory group could have a meaningful discussion regarding the application of successful classification. This classification study has been the first phase in what is expected to be a continuing effort that will span several years and investigate sites with increasing challenges.

1.4 ABOUT THIS REPORT

This report is intended to provide an overview of the key results of this pilot program for project managers, regulators and contractors. The focus of this report is on commercial instruments and available processing. Because of their dramatic success, we make a few observations about the potential of specialized emerging sensors. However, the material covered in this report represents only a small part of a much larger study, and notably absent here is any discussion of advanced and innovative processing techniques. More information about the entire demonstration and these topics in particular may be found in the individual demonstrator reports (Refs. 3–6) and an independent performance assessment by the Institute for Defense Analyses. (Ref. 7)

We begin with a description of the site and an overview of the program approach. We then describe the detection and classification performance. As a key factor in successful analysis is the data quality, a section is devoted to this subject. This is followed by a discussion of costs and a summary of the program conclusions. Finally, we consider how classification could be implemented on a real site.

2 FORMER CAMP SIBERT

2.1 TEST SITE OBJECTIVES

The former Camp Sibert, Alabama was selected for this study. The site has generally flat terrain, limited vegetation, limited to moderate geologic interference, a simple clutter environment, and most importantly a single munitions type. The demonstration area (Figures 2-1 and 2-2) was selected on a section of a known target a sufficient distance from the target center where the average anomaly density was low enough to allow for classification of individual items.

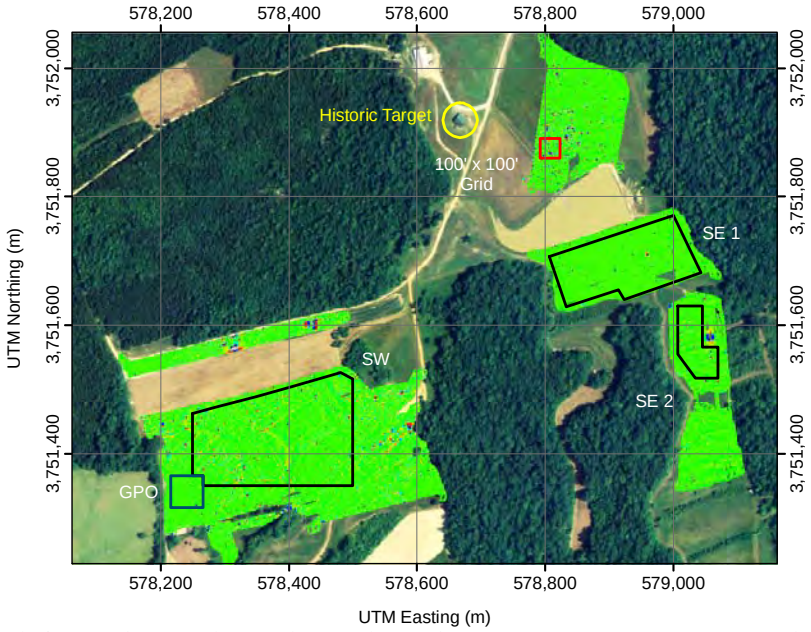

Figure 2-1. The former Camp Sibert Classification study area located on Site 18.

Figure 2-2. Former Camp Sibert Classification study area.

2.2 SITE HISTORY AND CHARACTERISTICS

The former Camp Sibert consisted of 37,035 acres in Etowah and St. Clair counties acquired by the U.S. Army in July 1942. Historical records and investigation of the site indicate that Camp Sibert was

used extensively as the main training camp for chemical warfare troops during World War II. As part of this training, troops fired large quantities of conventional and chemical munitions and handled live chemical agents during their training in chemical storage, weapons filling, and decontamination of equipment. During the investigation of historical records, numerous areas were identified as possible contaminated sites. (Ref. 8)

The camp was closed at the end of the war in 1945, and the chemical school transferred to Ft. McClellan, Alabama. The Army declared the property excess and transferred it to the War Assets Administration on 18 November 1946, and then to the Farm Mortgage Corporation. The government terminated the leases on the area on 13 December 1946. After decontamination of various ranges and toxic areas in 1948, the land was transferred to private ownership. The airfield was transferred to the City of Gadsden and is now the Gadsden Municipal Airport. The majority of the property has been privately owned since 1949 and either farmed or left as woodlands.

The Classification Study site is located within the confines of Site #18, Japanese Pillbox Area No. 2. Simulated pillbox fortifications were attacked first with white phosphorous 4.2-inch chemical mortars followed by troop advance and another volley of high explosive-filled 4.2-inch mortars. Assault troops would then attack the pillboxes using machine guns, flamethrowers, and grenades. There is historical evidence of intact 4.2-inch mortars and associated debris at the site. A limited geophysical survey of Site 18 was conducted prior to this study as part of the ongoing munitions response process and multiple anomalies were identified. (Ref. 9)

2.3 STUDY SITE OVERVIEW

The Classification Study site was chosen to be outside of the high density target center. Three areas (SW, SE 1, and SE 2) totaling 15 acres were selected to make up the final study area as seen in Figure 2-1 and their locations in relation to the historical target are noted. The locations of these areas were determined from results of a pre-demonstration magnetometer survey discussed below.

2.4 DEMONSTRATION PREPARATION

Several activities occurred prior to data collection to ensure the resulting data would support a successful demonstration. These activities include an initial magnetometer survey, excavation of a 100 foot × 100 foot site characterization grid, construction of a geophysical prove out (GPO), and emplacement of seed targets.

2.4.1 Initial Magnetometer Survey

A 100% coverage magnetometer array survey was conducted on about 50-acres of Site 18 to guide selection of the final 15-acre demonstration study area. The main objective was to select a study area with anomaly density between 100 and 200 anomalies per acre. This density would provide a sufficient number of targets to support statistical analysis of the results, while still allowing for successful classification of isolated targets. Three noncontiguous areas were chosen to make up the demonstration site. The two southeast areas (SE 1 and SE 2) were found to be within the desired density range. The density in SW was above this range due to local geology.

In addition, the initial magnetometer survey was used to guide selection of locations for the GPO, the 100 foot × 100 foot site characterization grid, and the seed targets.

2.4.2 Site Characterization Grid

A 100 foot × 100 foot site characterization grid was excavated to provide information about the types and depths of munitions and clutter on the site. The grid was selected near the target center where anomaly densities were higher to provide the maximum information about the items on the site. A total of 302 anomalies identified in the initial magnetometer data were dug, and information regarding the identification and depth of the recovered objects was provided as training data to the demonstrators that were applying classification approaches. The items were separated into classes shown in Table 2-1, and examples of the excavated items are shown in Figure 2-3. All the recovered munitions debris was associated with the 4.2-inch mortar, and no evidence was found to indicate that any other munitions types were present. It was intended that the depth information from the recovered items would be used to help guide the depth distribution of the seeded items. However, due to schedule constraints, their excavation was not completed prior to seeding the study areas, and only one intact mortar was ultimately recovered.

Table 2-1. Class of items excavated from the site characterization grid.

Class	Number
Intact Mortar	1
Munitions Debris	134
Cultural Debris	3
Hot Soil/No Contact	164

Figure 2-3. Items recovered during the excavation of the site characterization grid. A baseplate is on the left and an intact mortar is on the right.

2.4.3 Geophysical Prove Out

A GPO was established to verify detection thresholds for all instruments. The intent of the GPO was to verify that the targets of interest were detected at the depths of interest at the selected threshold under site-specific conditions. For each sensor, the threshold used to select the target list on the demonstration site was verified using the GPO results.

The location of the 50 m × 50 m GPO was chosen from the initial magnetometer survey. All items detected in the initial magnetometer survey of this area were excavated, and then the area was re-surveyed using an EM-61 cart, and any new detections were also excavated. The area was then

seeded with inert 4.2-inch mortars, the munitions known to have been used on this site. The seeded mortars used in the study were recovered from a site in Montana. The GPO contained 38 targets, distributed as shown in Table 2-2: 30 were inert 4.2-inch mortars and 8 were splayed half-rounds (recovered at Camp Sibert), which were expected to be a common clutter item. The locations and depths of these targets were unknown to the data collectors. The burial depths were biased shallow to provide high signal-to-noise training data for classification teams. A few rounds are buried to a depth of 11 times their diameter, the *de facto* expectation for detectability with modern geophysical equipment, to verify detection by the geophysical sensors. Targets were separated by a minimum of 6 m in all directions, and the target locations were not on a regular grid.

Table 2-2. GPO targets.

Munition type	Quantity	Depth Range	Orientations
4.2-inch inert mortar rounds	30	Flush buried to 1.17 m	Random orientations, predominantly from ±45° from horizontal, few vertical targets
4.2-inch mortar splayed, half-rounds	8	0.1 to 0.43 m	Either flat or on edge

2.4.4 Test Pit Measurements

A 1-m deep test pit was dug near the site for the use of the demonstrators and an example 4.2-inch mortar provided. Prior to beginning survey operations, each of the data collection demonstrators used the pit to collect signatures from the mortar at a variety of depths and orientations. These data were used to establish the detection thresholds for each sensor and as training data for the analysis demonstrators.

2.4.5 Seeding the Survey Area

The demonstration area was seeded with 151 recovered inert 4.2-inch mortars. Two of the 151 seed mortars were buried in locations where they formed a cluster with either an existing clutter object or a geological return. These two mortars were removed from all subsequent analysis, as were all clusters with overlapping signatures. The scoring throughout was done using the 149 isolated mortars. The locations and depths of these targets were unknown to the demonstrators. The exact (*x, y*) location, depth to the center of the target, and orientation were recorded for each emplaced item. Objects were not emplaced at depths in excess of 11 times their diameter, 1.17 m for the 4.2-inch mortar rounds. The emplacement distribution of these inert rounds is summarized in Table 2-3. Only *in situ* clutter was used in this study, and no additional cultural clutter, munitions-related scrap, or geology was seeded.

Table 2-3. Blind seeded targets.

Munition type	Quantity	Depth Range	Orientations
4.2-inch inert mortar rounds	151	Flush buried to 1.17 m	Random orientations, predominantly from ±45° from horizontal, few vertical targets

3 PROGRAM DESIGN

3.1 OVERALL APPROACH

The objective of the study was to evaluate classification, as opposed to detection. Multiple classification approaches were applied to data collected using seven different sensor platforms. For comparisons of different classification approaches to be straightforward, a common set of detections for each data set was required. The detection stage was done by a program office team that was separate from the data processors. The approach to detection is described below. For each data set, a common list was passed to all of the classification demonstrators set to attempt classification. The classifiers were required to make a declaration for each anomaly detected in each data set that they elected to analyze.

All the targets on the detection lists were dug and assigned ground-truth labels. These labeled data, including the seeded targets, were segregated into training and testing data. All the truth information for the training data was provided to the processors and used to train their algorithms. The truth labels for the remaining data were sequestered, and these were used for blind testing. The processors were required to provide their assessment of the munitions/clutter labels for each item in the test data part of the detection list. The labels were compared to truth by an independent third party to score performance.

3.2 EVALUATING SUCCESS

The main goal for classification in the pilot program was to identify items that were NOT MUNITIONS with high confidence. For stakeholders to find it acceptable to treat a detected anomaly as something other than a potential munitions item—that is, to forego safety measures or leave an item unexcavated—these determinations must be very certain. Thus, for the purposes of this demonstration, the main failure was misclassifying a target of interest (4.2-inch mortar) as a nonhazardous item. This is commonly termed a false negative.

Evaluating the performance in the demonstration required that the objects uncovered be divided into TARGETS OF INTEREST and CLUTTER.

TARGETS OF INTEREST: The only munitions type expected on the site was the 4.2-inch mortar. Since it is not possible to discriminate explosive from practice rounds using the methods demonstrated and since a practice round would concern the public in the same way that an intact high-explosive round would, a TARGET OF INTEREST for the purposes of the pilot program was defined to include:

- Intact 4.2-inch mortars, both high explosive and practice;
- Sizeable pieces of the mortars, which would be sufficiently munitions-like to alarm the public; and
- Items with a geophysical signature that is indistinguishable from munitions (i.e., steel pipes of similar size).

During the demonstration, no intact munitions were found and no pipe-like items were found. The largest pieces of munitions found were so-called half rounds, which were large, flattened objects

formed when the mortars split along their long axes. These were not at all munitions-like in their appearance.

CLUTTER: Clutter items included fragments, splayed half rounds, single fins, nose cones, base plates, cultural debris, and geology.

Prior to the demonstration, it was decided that hazardous components of munitions, such as fuzes, spotting charges, and bursters, while not the main focus of the demonstration, would be treated in a separate post-demonstration analysis. Ultimately, none of these component items were found, and this aspect of the classification problem could not be explored. Similarly, it was also unknown whether any grenades would be found. The historical records indicated they had possibly been used, but no physical evidence had been seen. Neither grenades nor grenade fragments were found in the classification grid, nor in all the ground-truth items dug, so the only item of interest remained the 4.2-inch mortar.

3.3 DATA COLLECTION

The classification pilot study consisted of several combinations of data-collection platforms and analysis approaches, ranging from careful application of commercial survey instruments to a prototype system specially designed to maximize detection and classification of munitions. The commercial survey systems were deployed to collect data on 100% of the site, called SURVEY mode. Three sensors were deployed to collect high-density data over selected individual anomalies, detected by other sensors, called CUED mode. Data-collection plans were generated by all data collectors and shared with the data processors prior to deployment.

In survey mode, the following sensors covered 100% of the site. Data were acquired by running a sensor in closely spaced lines, similar to the pattern of a lawnmower cutting grass. Care was taken when designing the data-collection protocols to ensure that data of a sufficient quality to support advanced analyses would result. For the most part, this involved controlling data density and system noise.

- **EM61-MK2 CART:** The on-site contractor, Parsons, used a standard cart platform EM61-MK2 system. Typical industry-standard centimeter-level accuracy Global Positioning System (GPS) equipment was used for geolocation and navigation. The survey lane spacing was specified as 0.5 m, compared to the 1 m lane spacing used on the production work on other parts of Camp Sibert. The sensor height above ground was the standard 40 cm. Figure 3-1a shows this system, which will be referred to throughout the report as EM61-MK2 CART. (Ref. 10)
- **Multisensor Towed Array Detection System (MTADS):** Three towed-array systems from the MTADS family developed by the Naval Research Laboratory (NRL) were used. For all, data were taken in the MTADS standard configuration using centimeter-level accuracy GPS and an Inertial Measurement Unit (IMU) for geolocation and platform orientation. (Ref. 11)
    - **EM61-MK2 array:** The MTADS EM system uses an array of three overlapping 1 m^2 EM61-MK2 sensors that have been modified to increase the transmit current and adjust the receiver time gates from the standard sensor. The three sensors are pulsed and sampled simultaneously. Data are collected with 0.5-m line spacing and the array rides 33.5 cm above the

ground surface. Figure 3-1b shows this system, which is referred to throughout the report at EM61-MK2 ARRAY.

- **Magnetometer array**: The MTADS magnetometer (MAG ARRAY) platform houses a 2-m wide array of eight G-822A Cs vapor magnetometers, spaced 25 cm apart and 25 cm off the ground.
- **GEMTADS:** The NRL GEMTADS incorporates an array of three frequency domain 96-cm diameter GEM-3 sensors arranged with two side by side and one centered and offset to the rear, for an effective side-to-side spacing of 0.5 m. The sensors have been modified to increase transmit current and are sampled sequentially. Data are collected at 33.5 cm height above ground.

Figure 3-2 shows an example of the EM61-MK2 ARRAY data from the GPO. The locations of the seeded mortars and half shells are marked over the data.

(a) (b)

Figure 3-1. Panel (a) shows the EM61-MK2 CART system and Panel (b) the EM61-MK2 ARRAY system.

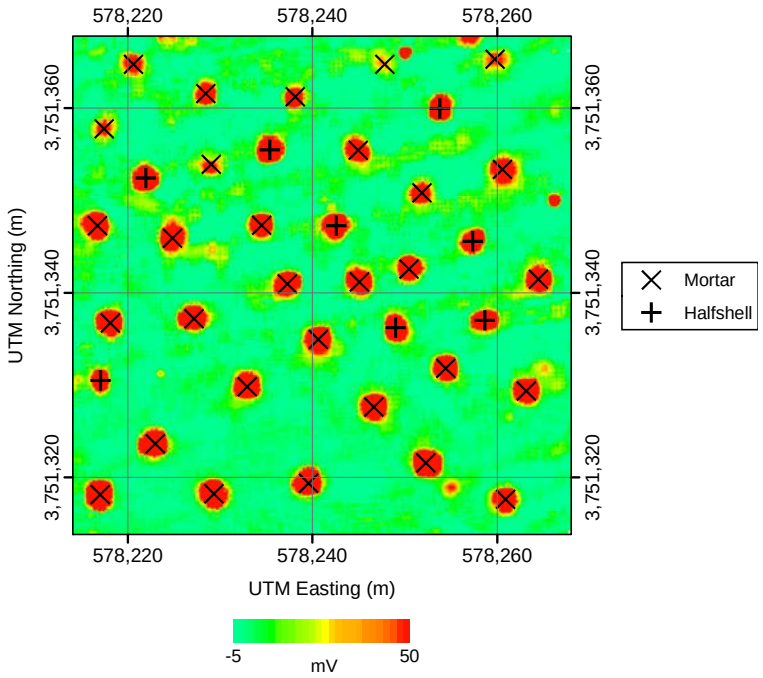

Figure 3-2. EM61-MK2 array data over the GPO.
The locations of the seeded mortars and half shells are marked.

In a cued mode, the following sensors were used to collect high-density data on approximately 200 anomalies selected from the survey data. Targets were selected to capture a representative sample of fit quality, sizes, and depths as estimated from the survey data. Both sensors gathered data on the same set of targets.

- **EM63:** The EM63 is a 26-channel time-domain EM instrument. It was deployed on an air-suspension cart to gather high-density cued data in a precise pattern centered over the selected targets. It used Robotic Total Station (laser) for positioning and an IMU for orientation. (Ref. 12) Figure 3-3 shows the cued EM63 data collection.

Figure 3-3. Cued EM63 data collection.

- **Hand-Held GEM-**3: This frequency-domain EM sensor was used in cued mode to gather data over grid of static points for each anomaly. A template was used for positioning the data collection locations.

The Lawrence Berkeley National Laboratory (LBNL) UXO Discriminator (BUD) operated in both a survey and cued mode:

- **BUD:** This developmental EM cart system was designed to collect sufficient data to fully characterize the EM signature from a single measurement location. It is composed of three orthogonal transmitters for target illumination, and eight pairs of differenced receivers for response recording. It measures the entire decay curve up to 1.2 ms after the transmitters are turned off. It surveyed about 5 acres of the site in survey mode and also collected data on all the cued targets in the remainder of the site. A photo and schematic of BUD are shown in Figure 3-4 (Ref. 6)

MAG AND FLAG: A mag-and-flag survey was conducted on a portion of the site to provide a point of comparison. The on-site contractor performed a typical operation with instructions to detect all of the targets of interest. (Ref. 10) All flag locations were dug and scored as though they had been declared potential targets of interest by the operators. No follow-on classification was possible.

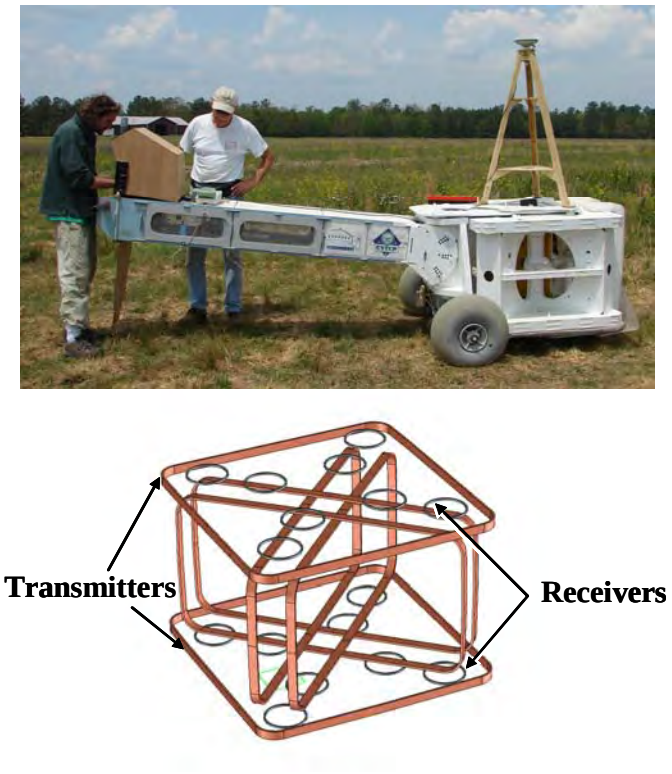

Figure 3-4. Photo and schematic of BUD.

3.4 CLASSIFICATION APPROACHES

Four groups demonstrated processing approaches. The basic classification method for all the demonstrators involved using a geophysical model to estimate target parameters that may be useful in making a classification decision. For all the sensors except BUD, this process involves using data from multiple spatially diverse locations that together fully characterize the signature. An example of a small section of field data encompassing an anomaly, called a data chip, is shown on the left panel of Figure 3-5. During the processing, the field data are used to extract the values of the model parameters. The right panel shows the modeled chip, which depicts the anomaly as it is predicted using the best fitted parameter values. When meaningful parameter values are arrived at, the two should look substantially similar.

BUD collects sufficient data at a single spatial location to support model-based parameter estimation. Although the processing approaches differ in their manner of implementation, all are based on a dipole model.

Some of the parameters that were considered included:

- The magnetic dipole moment, which is related to the physical size of the object;
- The magnetic polarizability, which relates to the object's size and aspect ratio; and
- The electromagnetic decay constant, which relates to the object's material properties and wall thickness.

Here the estimated size of the object should not be confused with the spatial size or footprint of the anomaly. While it is true that large, deep objects will give rise to anomalies with a greater spatial

dimension than small, shallow objects that may have comparable amplitudes, anomaly size is not a rigorous, direct substitute for object size.

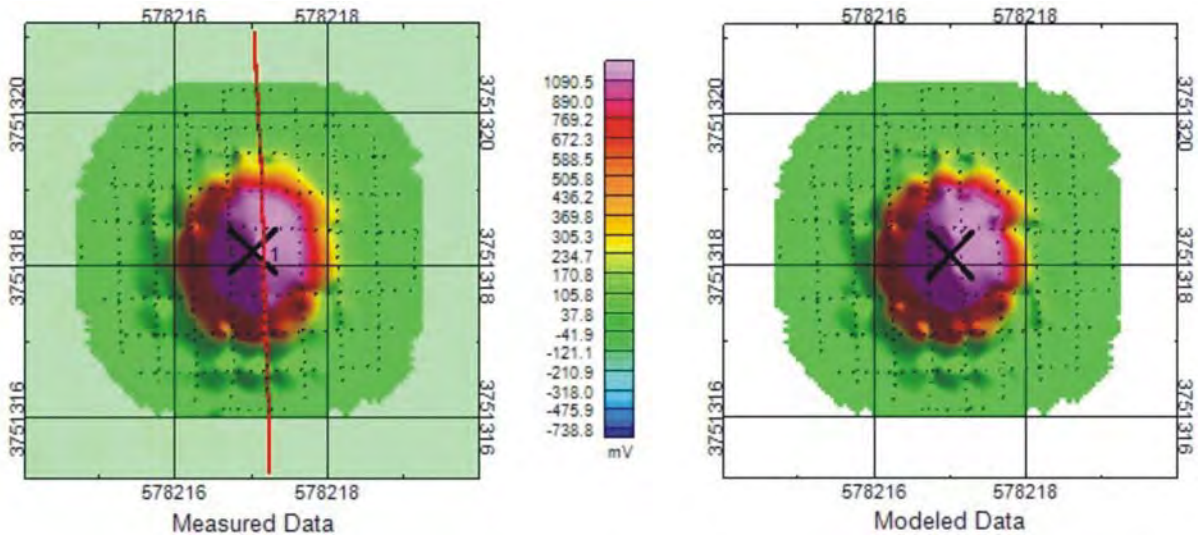

Figure 3-5. Example of a measured EM61-MK2 data chip of an anomaly (left) and a model result (right). Axes' units are in meters.

The basic flow of the classification approaches was the same for all demonstrators and is summarized in the flow chart in Figure 3-6. The classification demonstrators began with target lists provided by the program office. These lists contained all the anomalies detected by each sensor, as described in Section 4. Each anomaly was analyzed by the processing teams to extract parameters by fitting the data to a model. Inadequacies in the model, noise in the data, or difficulty in the mathematical process used to fit multiple parameters to the measured data will result in variation in these parameter estimates. Sometimes noisy data or a model insufficiency will yield a result that is nonsensical or will cause the estimation process to fail to converge on an answer at all. Although the demonstrators were requested to provide estimated parameters for each target analyzed, in some cases where meaningful fits could be not be obtained, items were identified as "Can't Analyze." Since no classification decision can be made, all items in this category must be treated as potential munitions.

Some of the truth data was provided to each demonstrator for algorithm training, so that the parameters that were most useful for classification could be determined. It is expected that some parameters will be more useful than others: on this site, with large munitions of interest, it was expected that the size-related parameters would be particularly useful.

Once the parameters are estimated, a mechanism is needed to decide whether the corresponding object is a target of interest or not. Several types of classification processing schemes were evaluated in the classification study. These included both

- Statistical classification: Computer algorithms evaluate the contributions of each parameter to defining munitions likeness based on "training" on a subset of the data for which the identities of the objects are known. Then the unknown objects are prioritized based on whether their parameters are statistically similar to known objects in the training data.

- Rule-based classification: A data analyst inspects the training data and the associated parameters to make a "rule" about how unknown objects will be sorted. For example, a rule may be defined so that all objects are sorted based on their "size" and decay constant.

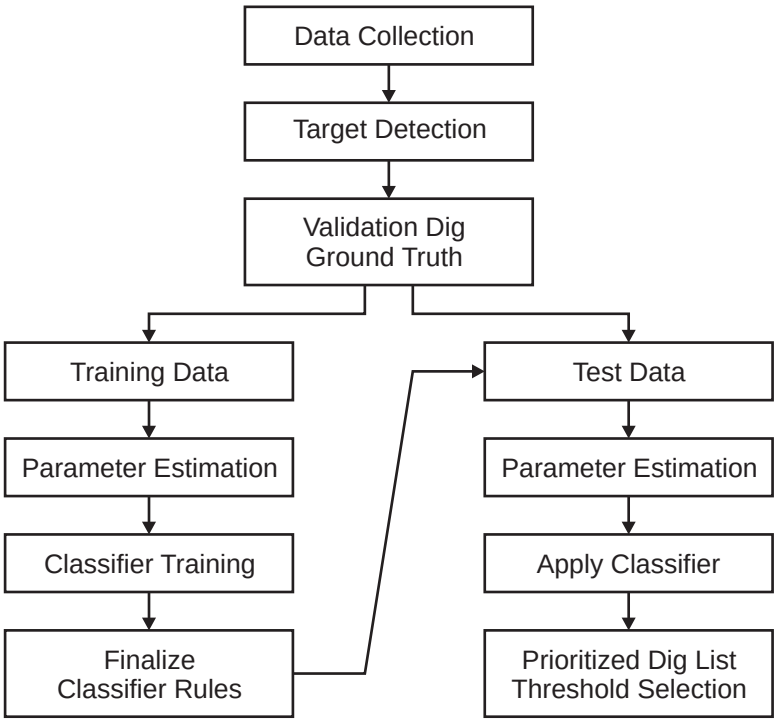

Figure 3-6. Work flow of classification demonstrations.

The final step in classification is delineating the items of interest from those that are not. For example, in the case of a statistical classifier, all the anomalies are ordered by the likelihood that they belong to the class of the targets of interest. These likelihood values do not represent a yes/no answer, but rather a continuum within which a dividing line or threshold must be specified. Depending on the application, this threshold may be set to try to avoid false positives, which may come at the expense of missing some items of interest, or it may be set to try to avoid false negatives, which will come at the expense of a greater number of items not of interest. In this program, where missing an item of interest represented the most serious failure, demonstrators selected thresholds to try to retain all the detected munitions.

Table 3.1 provides a summary of the classification approaches demonstrated. Those in bold are discussed in this report. The focus of this report is on commercial instruments and available processing that were successfully demonstrated and could be implemented today. Because of their dramatic success, we make a few observations about the potential of specialized emerging sensors. However, the material covered in this report represents only a small part of a much larger study, and notably absent here is any discussion of advanced and innovative processing techniques. Details of each approach and all of the results can be found in the individual demonstrator reports. (Refs. 3–6)

3.5 GROUND TRUTH

All targets detected by all sensors were dug up and identified to provide truth data. A master excavation list was produced by the Program Office from the union of the anomalies from the three

survey data sets (EM61-MK2 ARRAY, MAG ARRAY and GEMTADS). Duplicate detections were consolidated. Additional anomalies from the EM61-MK2 CART, BUD, and MAG&FLAG surveys were added. Clusters of overlapping anomalies were dug to provide data for future advanced processing research, but these data were not used for evaluating the performance in this study. No attempt was made to do classification on targets with overlapping signatures. For each anomaly, the location, depth, and description were recorded and a digital photo was taken. Each excavated item was then assigned to a class of either target of interest or clutter.

Table 3-1. Data-processing approaches demonstrated in the pilot program.[1]

Demonstrator	Data Analyzed	Processing Approach Summary
LBNL	**BUD**	• **Custom software for extraction of polarizabilities for the entire measured EM signal decay** • **Rules-based classification: library match to 4.2-inch mortar**
SAIC, Inc.	**EM61-MK2 Cart** **EM61-MK2 Array** **MAG Array** GEM Array	• **UX Analyze beta version: routines transitioning to Geosoft Oasis for fitting MAG and EM61-MK2 data** • **Statistical classifier** • Custom software for improved parameter extraction • GEM custom analysis and library match
Sky Research	**EM61-MK2 Cart** **EM61-MK2 Array** **MAG Array** EM63 Cued	• **UXO Lab: collection of geophysics analysis software for fitting MAG and EM61-MK2 data** • **Rules-based classifier based on size and decay constant** • UXO Lab for EM63 data • Cooperative inversion of MAG and EM data combinations • Statistical Classifier
Signal Innovations Group	EM61-MK2 Cart EM61-MK2 Array MAG Array GEM Array EM63 Cued	• Custom software for extraction of parameters • Multiple statistical classifiers, supervised and semi-supervised • Traditional and Active Learning Training
Parsons	**EM61-MK2 Cart**	• **UXAnalyze for parameter extraction** • **Decision rule based on size and fit quality**

[1]The items in bold are discussed in detail in this report.

3.6 CLASSIFICATION PRODUCT

Demonstrators were asked to produce a ranked dig list for each sensor and processing combination as their primary product. These lists were constructed as shown in Table 3.2

Table 3.2. Model of Ranked Dig List.

Rank	Comment
1	
2	High confidence NOT munition
3	
	THRESHOLD
…	
…	Can't make a decision – clutter like
…	
…	
…	Can't make a decision – munitions like
…	
…	
…	High confidence munitions
…	
N-2	
N-1	Can't extract reliable features
N	

- GREEN: The top item in the list was that which the demonstrator was most certain does NOT correspond to a munitions item.
- RED: The bottom items were those that the demonstrator was most certain are munitions.
- YELLOW: A band was specified indicating the targets where the data can be fit in a meaningful way, but the derived parameters do not permit a conclusion.
- GRAY: Targets where the signal-to-noise ratio (SNR), data quality, or other factors prevent any meaningful analysis were deemed "can't analyze" and appended to the bottom of the list.
- THRESHOLD: A threshold was set at the point beyond which the demonstrator would recommend all anomalies be treated as munitions, either because they are determined to be so with high confidence or because a high-confidence determination that they are not munitions cannot be made. This is indicated by the heavy black dashed line.

3.7 SCORING METHODS

The demonstration was scored based on the demonstrator's ability to eliminate nonhazardous items while retaining all detected munitions. A common way to evaluate performance of detection and classification is the receiver operating characteristic (ROC) curve. An example in shown in Figure 3-7. The ROC curve is a plot of the probability of detecting and correctly classifying the munitions items versus the number of false positives (clutter targets). A perfect detector and classifier would detect 100% of the munitions and no clutter.

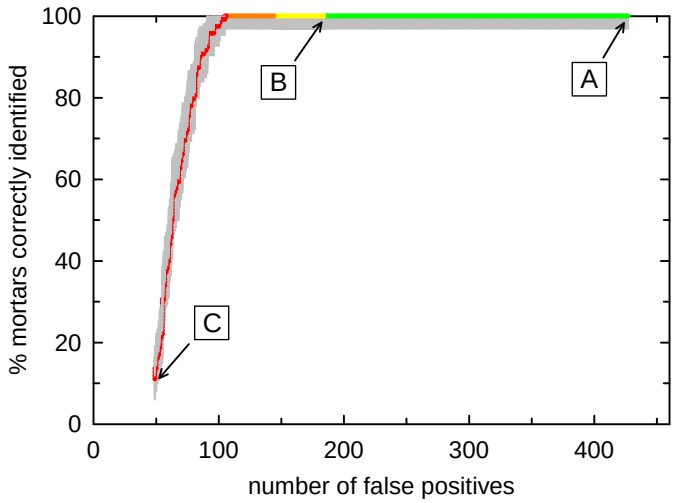

Figure 3-7. Example receiver operating characteristic curve.

The key regions to interpret the ROC curves used in this program are:

- A: In the absence of any classification, this sensor detected all the munitions and had more than 450 clutter items in the detection list.
- B: Based on classification, this is the demonstrator's threshold for the dividing point between munitions and not munitions. This demonstrator correctly identified all of the munitions and about 185 clutter items remained after classification.
- C: Targets to the left of this point were categorized as can't analyze and would need to be treated as potential munitions because no meaningful classification could be done. In this example, about 50 of the can't analyze targets were false positives, reflected in where the curve starts in the horizontal axis at about 50. Fourteen munitions (of the 149 munitions seeded) were also included in the can't analyze list, reflected in the curve beginning at 10% on the vertical axis.

4 DETECTION

4.1 ANOMALY–SELECTION THRESHOLD

The anomaly-selection phase of the demonstration was handled by the ESTCP Program Office team for all sensors except the BUD. To allow for comparisons of the classification approaches that were not confused by differences in setting detection thresholds, all processing demonstrators were asked to classify a common set of targets for each sensor. A detection list was generated by recording all locations for which the sensor signal exceeded a sensor-specific threshold. Since these individual sensor detection lists were the basis for all subsequent analyses in the demonstration, the detection threshold was set using a rigorous process.

The target of interest in this demonstration is a 4.2-inch mortar. Using the rule of thumb that an item is expected to be detected routinely at a depth equal to 11 times its diameter, the depth of interest for this was set at 117 cm. The signal from each sensor can be predicted accurately for this item of interest as a function of depth, and the detection threshold set to ensure detection of all 4.2-inch mortars while leaving smaller targets off the list. This is illustrated for the magnetometers in Figure 4-1, where the predicted peak magnetic anomaly amplitude for a 4.2-inch mortar in its most favorable and least favorable orientation is plotted as a function of depth. For a magnetic measurement such as this, the most favorable orientation is when the target is aligned with the Earth's magnetic field and the least favorable is when the target is perpendicular to the field.

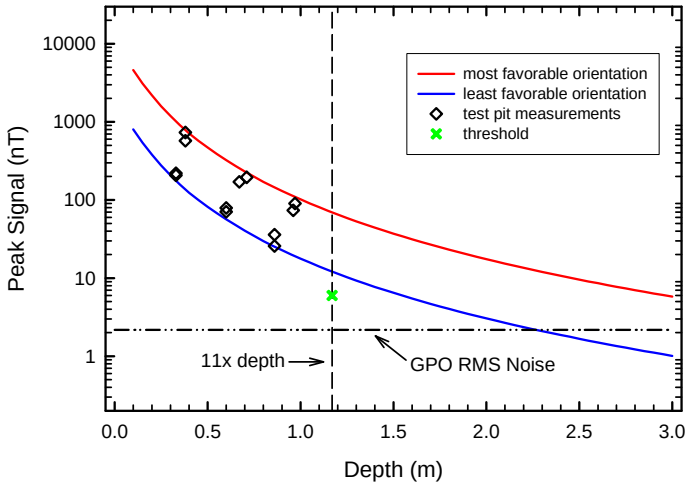

Figure 4-1. Predicted peak magnetic anomaly amplitude for a 4.2-inch mortar in its most favorable and least favorable orientations as a function of depth and measured values from a test pit adjacent to the GPO. Also shown are the system noise at the GPO and the detection threshold used for the magnetometer surveys.

To confirm the accuracy of these predicted anomaly amplitudes, the survey team made a number of measurements over the mortars at various orientations in a test pit adjacent to the GPO. The peak anomaly amplitudes from these measurements (black diamonds) are plotted in Figure 4-1, confirming the predictions. Based on these data, the smallest signal possible from the target of interest at the maximum depth can be confidently predicted and the detection threshold set accordingly. After consultation with the Program Advisory Group, the threshold for all sensors was set at 50% of the smallest expected signal to provide a safety margin. The actual threshold used for the magnetometer surveys is shown in green in Figure 4-1.

The final step in this process is to use the data from the GPO to confirm the threshold chosen. These data are from a blind (to the survey crew) survey of the GPO. The data from Figure 4-1 with these GPO results included are shown in Figure 4-2.

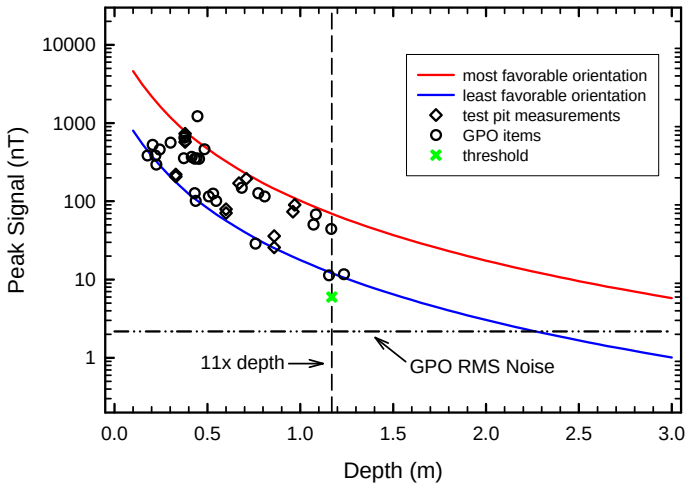

Figure 4-2. The data from Figure 4-1 plotted with the measured data from a survey of the GPO.

A similar process was used to set the threshold for the other survey sensors. The predicted signals for Gate 1 for the array of EM61-MK2, the most widely used geophysical sensor in munitions surveys, are shown in Figure 4-3. As before, the threshold is set at 50% of the smallest expected signal amplitude for the 4.2-inch mortar at a depth of 117 cm. For this sensor, the threshold is further from the system noise floor than was the case for the magnetometers sensors. Using the threshold plotted in Figure 4-3, there were 43 declarations in the GPO which correspond to detection of the 38 emplaced items (30 mortars and 8 half shells) and 5 false positives. The magnetometer survey resulted in 133 declarations in the GPO, substantially more false positives than the EM61-MK2 array.

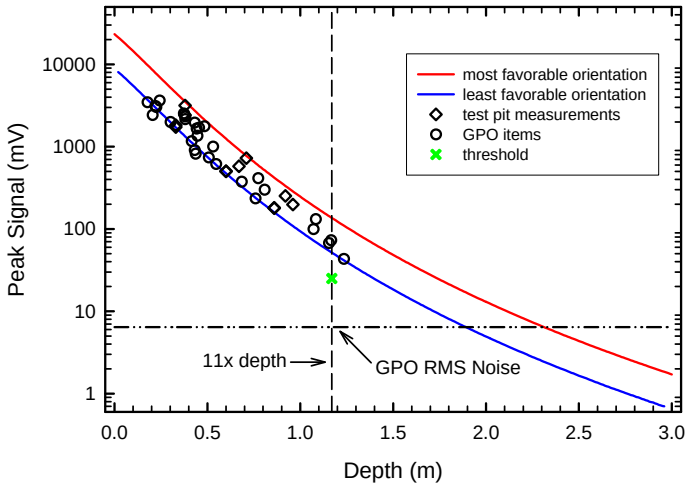

Figure 4-3. Predicted peak of the EM61-MK2 array anomaly amplitude in Gate 1 for a 4.2-inch mortar in its most favorable and least favorable orientations as a function of depth and measured values from a test pit adjacent and the GPO. Also shown are the system noise at the GPO and the detection threshold used for the EM61-MK2 array surveys.

Setting the threshold based on the signal expected from the item of interest makes an important contribution to the classification problem. At many sites, the detection threshold has been set at some multiple of total system noise, irrespective of item of interest; often this multiple has been 1.5. Figure 4-4 shows the number of detections in the EM61-MK2 ARRAY data as a function of the detection threshold. Lowering the detection threshold from the 25 mV determined from Figure 4-3 to 1.5 times the site noise would result in adding nearly 150% more locations to the target list. These additional targets are necessarily low signal-to-noise targets, which are often difficult to extract reliable features for and predominately end up in the "unable to analyze, must dig" category.

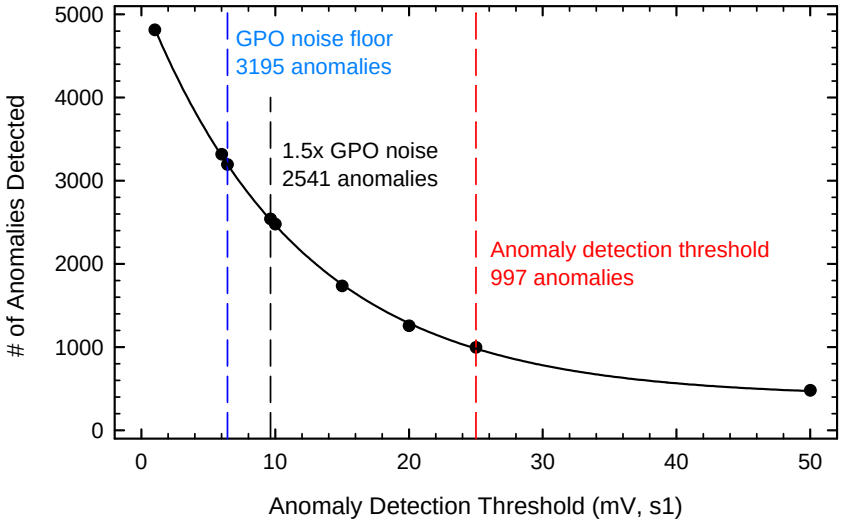

Figure 4-4. Number of automated anomaly detections as a function of detection threshold for EM61-MK2 ARRAY data. Thresholds corresponding to the system noise, 1.5 times the system noise, and the threshold used in this demonstration are marked.

4.2 MASTER ANOMALY LIST

Using this method, all sensors detected all 149 seed targets. The area surveyed by BUD contained only 66 seed targets, all of which were detected. The number of anomalies included on the anomaly list for each of the survey sensors is listed in Table 4-1 along with the detection threshold determined for that sensor. These anomaly counts are far smaller than the number of anomalies listed on each sensor's detection list. After these detection lists were received by Institute for Defense Analyses (IDA) personnel, the lists were combined, and only those anomalies separated from neighboring anomalies by at least 2 m were added to the master list to avoid overlapping signatures. Current approaches are not expected to perform effective classification when signatures of nearby anomalies overlap.

Overlapping anomalies presented a particular issue for the magnetometer data in the SW area. This area exhibits moderate geological response, which results in a large number of near-threshold detections due to geology. In this area, first detections that were contained in clusters of overlapping signals were removed. Then, the magnetometer and GEM lists were further pared down to include only locations that corresponded to EM detections.

Table 4-1. Threshold used and number of anomalies selected by the survey sensors.

Sensor	Anomaly Detection Threshold	Anomalies on Master List	Total Anomalies on Detection List	Seed Targets Detected
Magnetometer Array	6 nT	969	7314	149
EM61-MK2 Array	25 mV	870	1304	149
EM61-MK2 Cart	19 mV, sum of three gates	633	951	149*
GEM-3 Array	1.3 ppm, Q_{ave}	1039	3485	149
BUD (SE 1 area only)	See Ref. 6	244	362	66

* In the original detection list provided by the contractor, one seeded item did not have a corresponding target selected. Upon further examination of the data, an anomaly that exceeds the target selection threshold is associated with this seed. This anomaly fell on the boundary of the grid that the contractor used to manage the data processing and was omitted from the original list in an accounting error. It was subsequently added to the master list for inclusion in the classification step.

4.3 MAG AND FLAG DETECTIONS

The Mag & Flag team surveyed a 100 foot × 100 foot area in SE1. The team marked 49 anomalies in this small area. For comparison, the magnetometer array declared 39 anomalies in this same area, of which 27 made it onto the master anomaly list (the rest were removed as part of clusters). Both groups correctly marked the 4 mortars seeded in this small area but, of course, the Mag & Flag detections do not allow for further processing so all 49 detections must be dug. Of the 27 magnetometer array targets on the master anomaly list, 20 were correctly declared nonhazardous in the classification step so only 3 clutter items remained as false positives, and the remaining 4 were correctly classified seed targets. These results are summarized in Table 4-2.

Table 4-2. Detection and classification results in the Mag & Flag area.

	Magnetometer Array	Mag & Flag
Total Anomalies Declared	39	49
Number on Master List	27	49
Number of True Positives	4	4
Number of False Positives after Classification	3	45

5 CLASSIFICATION RESULTS

The data analysis occurred in three phases. In the preparatory phase, the demonstrators were given the appropriate survey data sets and the master anomaly list developed by the Program Office team. Each team then preprocessed the data as required, extracted a portion of the data corresponding to each of the anomalies to be classified, and submitted these data chips to their geophysical inversion routines to estimate target parameters. This led to the training phase of the data analysis demonstrations. The identities of approximately 200 of the anomaly sources were distributed to each demonstrator. These identities, along with the previously estimated target features, were used to train the demonstrators' classification algorithms and determine classification boundaries. The final phase was the blind testing of each classification algorithm. For each data set for which they were responsible, each demonstrator prepared one or more prioritized anomaly lists. The anomalies were arranged from most confident that the item does not correspond to intact munitions to most confident that the item is an intact munition.

5.1 FEATURE EXTRACTION

To perform the feature-extraction step, the demonstrator must first select the data points associated with each anomaly. Those data are then submitted to feature extraction routines. For all the demonstrators in this program, these routines consist of a model-match to a physics-based model of target response. The features estimated by these procedures include target position, a rough estimate of target size, and an indication of goodness-of-fit from the magnetometer data and position, orientation, shape, and goodness-of-fit from the EM data. A histogram of the magnetometer-derived "size" parameter estimated by one of the demonstrators is shown in Figure 5-1. Here size refers to the estimated physical size of the object, as opposed to the size of the anomaly footprint.

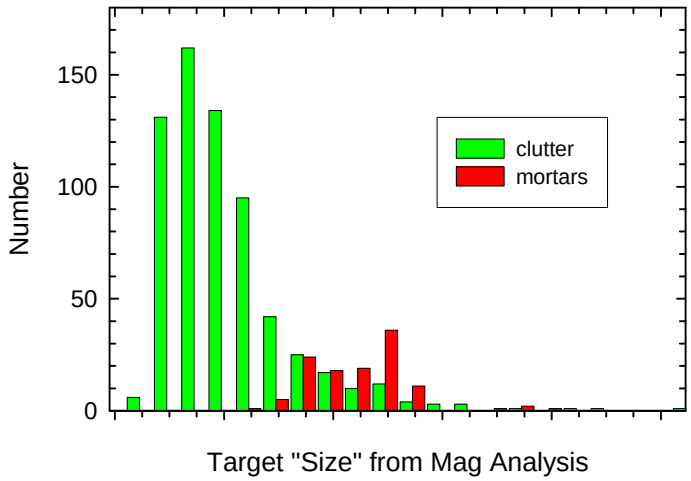

Figure 5-1. Histogram of estimated "size" extracted from the magnetometer array data for the 769 anomalies successfully analyzed. While related to target size, this surrogate feature does not directly correspond to a physical dimension, so the units and values of the *x*-axis are suppressed.

This demonstrator was able to extract reliable parameters for 769 of the 969 magnetometer anomalies on the master target list. Figure 5-1 shows that the distribution of "sizes" extracted from the clutter objects peaks at a value much smaller than the range of values observed for the 4.2-inch

mortar. This reflects the fact that the overwhelming majority of clutter items in the demonstration areas are smaller scrap and fragments.

A similar presentation of a related feature is shown in Figure 5-2, which plots the distribution of an EM "size" feature estimated by another demonstrator. In this case, the "size" feature corresponds to the log of the sum of the three principal axes responses of the target, which is related to target size. Similar to the magnetometer results, the clutter sizes peak at a value that is much smaller than the mortars, which cluster tightly together.

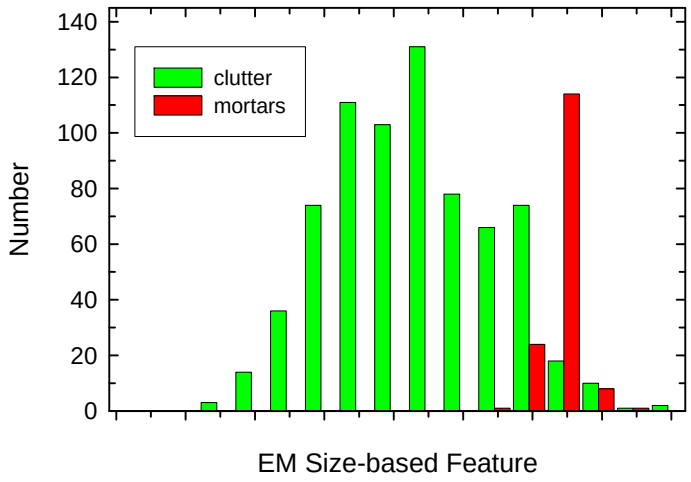

Figure 5-2. Histogram of an EM size-based feature for the 870 anomalies successfully analyzed from the EM61-MK2 array data set. While related to target size, this surrogate feature does not directly correspond to a physical dimension, so the units and values of the x-axis are suppressed.

Plots such as these can be made for each of the features extracted by each of the demonstrators. These estimated features have been compared in detail with the actual target characteristics by analysts at IDA, and these comparisons have been used to determine the relative effectiveness of each of the demonstrators' parameter estimation routines. Details of this analysis can be found in the IDA report (Ref. 7).

5.2 Algorithm Training

After reliable features have been extracted for each anomaly, the next step in the classification process is algorithm training. This can be as simple as setting a threshold for a classification method that relies on one target feature alone or as involved as training a statistical classifier on the distributions of multiple features associated with both munitions and clutter. All classification methods will perform better with some site-specific training data regardless of the complexity of the algorithms employed.

The identities of approximately 200 targets were given to the demonstrators for training purposes. These training targets were roughly evenly distributed throughout the three survey areas and included some of the cued targets as well as some targets covered by the survey systems only. This method was chosen to approximate digging a small portion of the field for training purposes, as might be done on a real site, while avoiding biasing the training results due to sampling a too narrow slice of available anomalies.

An example of the results of this training is shown in Figure 5-3. The features plotted were estimated using the EM61-MK2 array data. Plotted on the *x*-axis is the scaled amplitude of the response from the largest axis of the target at the first time gate of the EM61-MK2. This feature is related to the physical size of the object. As expected, the 4.2-inch mortars, denoted by red plus signs, all appear on the right side of the plot, as they are larger than most items in the survey area. Plotted on the *y*-axis is the ratio of the response from gate three to that of gate one. This serves as an approximation of the decay rate, with larger values corresponding to slower decay of the induced signal. With a more capable instrument this could be measured more directly, but even this approximation of the decay rate has classification value. Again, all the mortars, red plus signs, are in the upper part of the distribution, meaning they have slower decay rate than most of the items in the field.

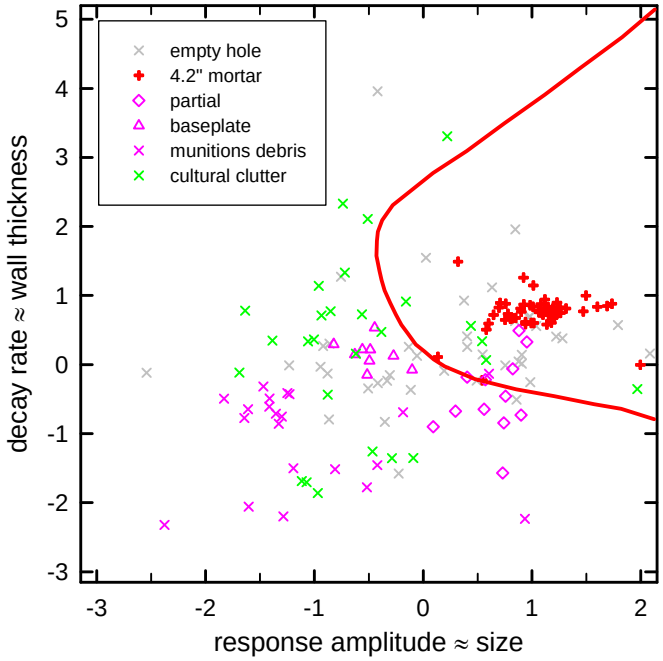

Figure 5-3. EM61-MK2 data used to train the classifier used by Sky Research for this demonstration. The identity of the source of each anomaly is denoted by the symbol used, and the lines represent the decision boundary in this two-feature space. The targets of interest are 4.2-inch mortars, and all other items are clutter. The features used by this classifier are discussed in the text.

This plot illustrates well the classification possible at this site. The mortars cluster in one region of the plot. A number of the partial rounds plot in the region of the munitions, but most of the fragments and junk are well separated from the 4.2-inch mortars on this plot. It is also noteworthy that several of the common clutter objects, such as partials and base plates, form clusters of their own, lending confidence that the parameters estimated are physically meaningful.

The line in Figure 5-3 shows the decision boundary developed using these data to train a statistical classifier. A measure of how munitions-like any particular target is can be estimated by its distance on this plot from the decision boundary and on which side it falls. By that measure, we can expect this classifier to perform well in the blind testing; the munitions are all contained in the "munitions" region of the plot, with the majority of the clutter outside that region and far from the boundary.

Similar plots were constructed by each demonstrator for each data set analyzed. These plots were used to tune the respective classifier parameters and decision boundaries. At the conclusion of this training phase, each demonstrator submitted a report to the Program Office describing its classification procedures and specifying the parameters and boundaries that would be used in the blind testing phase of the demonstration. After acceptance of these reports, each demonstrator proceeded to the blind test.

5.3 BLIND TEST RESULTS

In the blind test phase, each demonstrator submitted an anomaly list ranked from highest confidence not-munitions to highest confidence munitions for each data set analyzed, as shown in Table 3-2. The first item on the list, therefore, was the item that the demonstrator had the most confidence could be safely left in the ground. The individual anomaly lists were compared by IDA personnel to the ground-truth table developed by excavating each anomaly and carefully recording the result. (Ref. 7)

The results in this section are presented as ROC curves, which plot the percent of correctly classified munitions versus the number of false positives (i.e., clutter items). These curves are generated by stepping through the prioritized anomaly list item by item. For each entry, the percentage of munitions correctly classified and number of false positives are tallied, and the cumulative results are plotted. Ideally, all the high-priority munitions calls will correspond to actual munitions, resulting in 100% correct classification of munitions with a minimum number of false positives. These curves are generated to illustrate the tradeoffs between increasing detections and increasing false positives: correctly classifying more of the munitions comes at the expense of including more false positives. At the point where 100% of the munitions are correctly classified, the ROC curve can illustrate that digging more does not result in removing more munitions.

5.3.1 Magnetometer Data

The results of the analysis of the magnetometer data performed by Sky Research are shown in Figure 5-4. The individual points plotted on the curve are colored to correspond to the classification ranking assigned by the demonstrator and the threshold specified by the demonstrator is marked by the arrow. Several conclusions about this sensor/analysis combination can be seen from the figure, the most important of which is that of the 727 anomalies analyzed, this demonstrator was able to correctly declare 509 to be not-munitions with high confidence.

Working along the plot from left to right corresponds to looking at the anomaly list of Table 3-2 from the bottom to the top. This demonstrator was not able to estimate reliable features for 97 of the anomalies analyzed. These are placed at the bottom of the anomaly list and are always going to be treated as though they are potentially munitions. These anomalies are the reason the curve does not start at the bottom left corner of the plot. Of these 97, 2 corresponded to munitions and the other 95 were clutter. Of the 183 anomalies classified as high-confidence munitions, 117 were, in fact, munitions and 66 were clutter. These points are colored red in Figure 5-4. These anomalies capture all but 1 of the munitions while only resulting in 66 false positives. The remaining munition is 1 of the 13 anomalies classified as "unable to decide – munitions like." Including the anomalies classified as "unable to decide – clutter like" within the threshold of anomalies that must be treated as if they were munitions would only result in 22 more false positives.

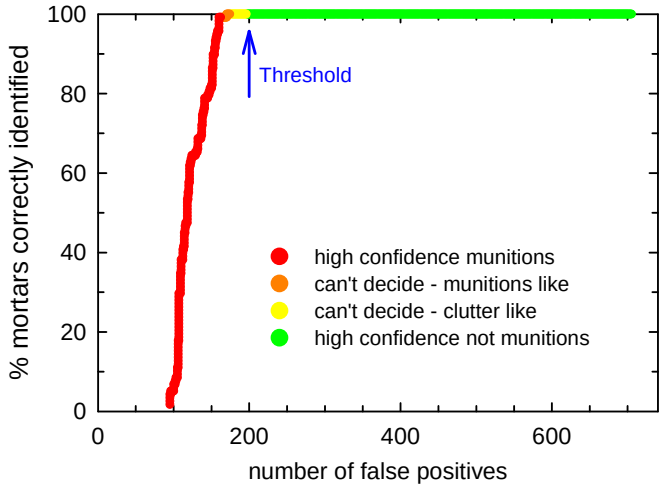

Figure 5-4. ROC curve resulting from the analysis of magnetometer features by Sky Research.

The sensor/analysis combination represented by Figure 5-4 should be judged very successful. More anomalies fell into the "unable to extract reliable features" category than would be optimum but, of the anomalies for which features were successfully estimated, the demonstrator was able to identify hazardous items with relatively high efficiency. The threshold was chosen appropriately. Only one mortar was classified as "can't decide – munitions like," and the number of items for which no decision was possible was small.

5.3.2 EM61-MK2 Data

The ROC curves for analysis of two of the EM61-MK2 data sets are shown Figure 5-5 This figure plots the results for analysis of EM61-MK2 data from both the cart and array using the UX-Analyze module of Oasis montaj. As can be seen from the figure, the results from the two platforms are virtually identical. In both cases, the analyst was unable to extract reliable features from just over 100 anomalies. After that, it takes just about 100 false positives to capture all 119 of the munitions in the test data sets. As was the case for the magnetometer data presented above, both of these analyses were able to correctly classify over 400 anomalies as resulting from nonhazardous items.

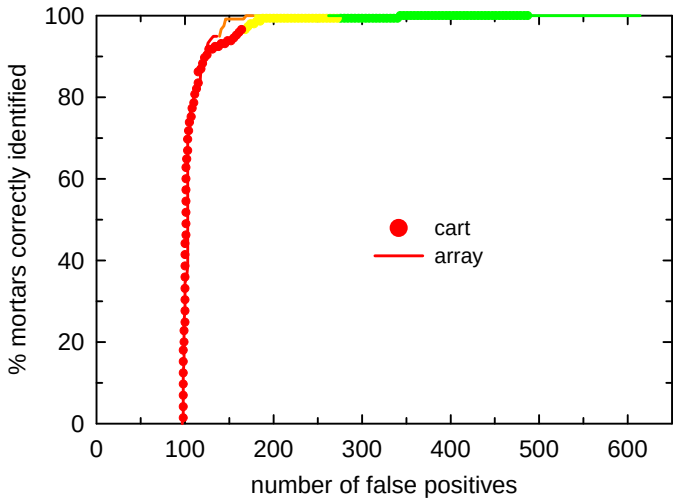

Figure 5-5. ROC curves resulting from analysis of EM61-MK2 cart and array data using UX-Analyze. The cart data were analyzed by Parsons and the array data by SAIC.

It is important to note that the cart results shown in Figure 5-5 were obtained using a commercial sensor and analysis with a freely available module of Oasis montaj. The cart data were collected using tighter lane spacing than would be standard for a detection-only survey (0.5-m spacing in this case), but otherwise, normal commercial data collection and operator procedures were used. Analysis of the EM61-MK2 cart data by either the developer of UX-Analyze or the contractor yield equivalent results, showing the ease of use and power of this module.

The threshold declared in conjunction with both analyses shown in Figure 5-5 was somewhat conservative; both anomaly lists included a significant number of anomalies corresponding to clutter after the last mortar was detected but before the threshold identifying high-confidence nonmunitions targets.

An informative aspect of the analyses is to compare the classification performance obtained using the size-based feature extracted from the EM61-MK2 CART data that was discussed in the previous section with that obtained using all available features in the classification. This would be equivalent to only using the data along the *x*-axis in Figure 5-3 for classification and ignoring the other available information. The comparative performance of these two methods is plotted in Figure 5-6. Although both classifiers correctly classified all mortars as "high confidence munitions," taking advantage of all available features allows the classifier to correctly recognize the mortars with fewer wasted false positives. This is seen in that the red filled circles rise much closer to vertical than the red line.

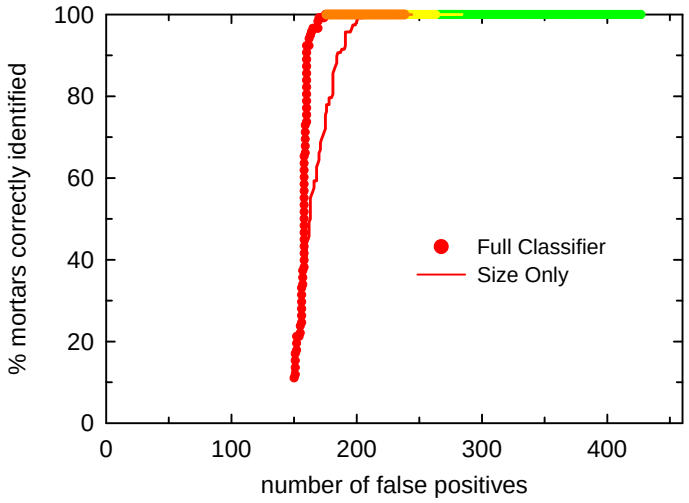

Figure 5-6. Comparison of the performance of a classifier based on size alone and with one based on all available features from the EM61-MK2 CART data. Analysis by Sky Research.

5.3.3 Cued EM63 Data

All three of the analyses presented so far have found that reliable features could not be extracted from a significant number of anomalies. In most cases, this is caused by insufficient signal-to-noise ratio to allow the geophysical inversion to converge. There are two primary sources of the noise that interferes with the analysis in surveys of this type, motion-induced sensor noise and apparent noise resulting from sensor location uncertainties. One way to lessen both of these noise sources is the use of cued investigations. Data can be collected in a small area around each anomaly on grid points or when moving slowly, lessening motion-induced noise and location uncertainty.

Figure 5-7 shows the results of a cued survey using the EM63 time-domain EM sensor. Unfortunately, this sensor/analysis combination also resulted in a significant fraction of anomalies for which reliable features could not be estimated. After these 32 anomalies classified as "unable to extract reliable features," the next 29 anomalies on the dig list (classified as high confidence munition) were munitions. This is reflected in the vertical red line in Figure 5-7. The threshold from this analysis was too aggressive; the final munition was not identified in this analysis until just into the anomalies classified as "high confidence not-munitions," but the top 68 anomalies on the prioritized list (more than half the total anomalies) were correctly classified as not hazardous.

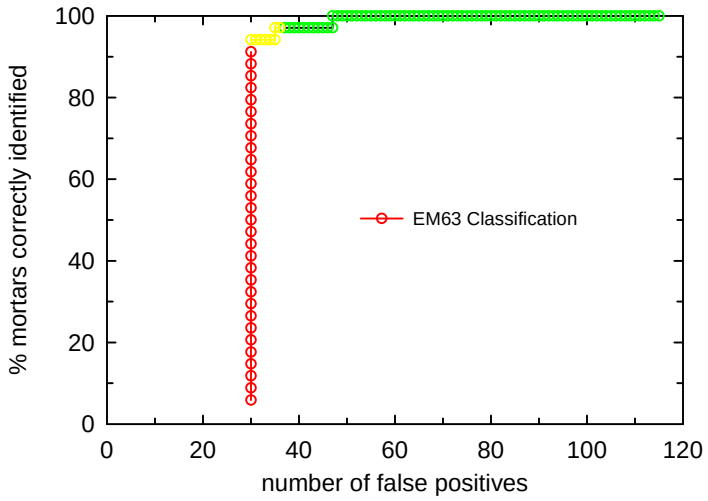

Figure 5-7. ROC curve resulting from the analysis of EM63 cued data by Sky Research.

Analysis of the inversion process for EM data for the anomalies classified as "unable to extract reliable features" often reveals that the data quality was insufficient to tightly constrain the depth estimated in the model-match procedure, and this unreliable depth estimate leads to unreliable estimates of response amplitudes. The data-quality requirements for an accurate depth estimate from magnetometer data are lower than for EM data, so one way to overcome this problem is to use the magnetometer-derived depth estimate to constrain the EM inversion. The results of classification using such a "cooperative" inversion for the EM63 cued data are shown in Figure 5-8.

Constraining the item depth using the magnetometer data has two obvious results. First, the number of anomalies classified as "unable to extract reliable features" is halved, a big improvement. Equally important, the quality of the inversions that result is much better, allowing the analyst to more accurately define the high-confidence clutter-classification threshold. For the cooperative inversion, all but one mortar are classified as "high confidence munitions," and no clutter is included in this category. The remaining mortar is the next item on the dig list; it was classified as "unable to decide-munitions like" and therefore fell on the correct side of the munitions/clutter threshold. For this analysis, the only false positives are the 16 anomalies for which features were not able to be extracted reliably.

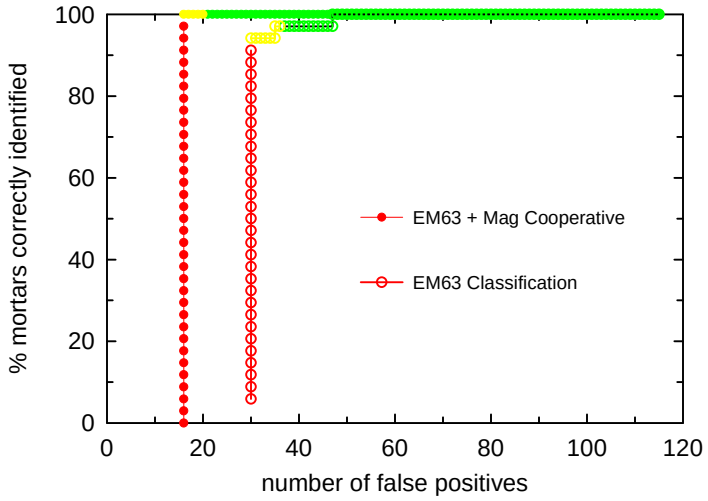

Figure 5-8. Comparison of the performance of a classifier using features extracted from the cued EM63 data only and one using features from inversions in which the depth was constrained by the depth derived from analysis of the magnetometer data. Analysis by Sky Research.

5.3.4 Advanced EM Sensor

As a final example, Figure 5-9 shows the results of the BUD survey of the SE1 area. This sensor is the first of a new generation of EM sensors with substantially more information content in the signals. It is still in the development phase so it was deployed as a cued sensor and surveyed only one of the sub-areas. Figure 5-9 represents a near-perfect ROC. The first 56 anomalies moving up from the bottom of the list were, in fact, munitions. The analyst was slightly too conservative specifying the threshold; the next six items on the list were marked as munitions but were clutter. After this, the 203 anomalies at the top of the list (over 75% of the total anomalies) were correctly marked as high confidence not-munitions.

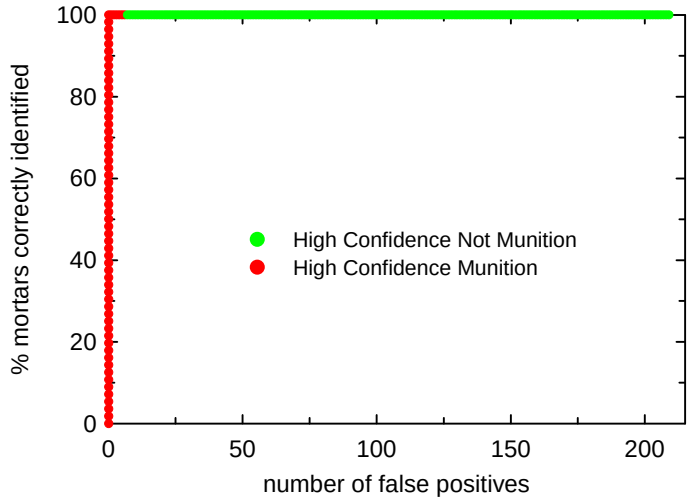

Figure 5-9. ROC curve resulting from the Berkeley UXO Discriminator survey of the SE1 area. Analysis by Lawrence Berkeley National Laboratory.

5.4 DETERMINATION OF THRESHOLD

In hindsight, we can examine how well the demonstrators did at setting the threshold. This is a critical aspect of classification. The demonstrator-specified location of the threshold for most analyses was remarkably accurate.

In many cases, the demonstrators were unnecessarily conservative in setting their threshold, including clutter items that could have been declared clutter. This is in line with the demonstration plan statement that the biggest failing is mislabeling a mortar as clutter. In a handful of cases, as illustrated by the cued EM63 analysis presented above, where one mortar is included in the high-confidence clutter category, the demonstrator was slightly too aggressive. In those cases, the addition of more information allowed the demonstrators to sharpen, and more correctly position, their decision boundary. The most capable sensor demonstrated in this program, the BUD, had a very accurate decision boundary as seen in Figure 5-9. Details on the threshold accuracy for all sensor/analysis combinations can be found in the IDA report. (Ref. 7)

6 DATA REQUIREMENTS

Classification requires data of higher quality and density than are typically collected for the detection of munitions. The primary concerns are with the data density, the geolocation precision of each data point, and the SNR. Numerous studies have shown that these factors can seriously degrade the ability to use model-based analysis to accurately extract target parameters. (Ref. 13) In this section, we will concentrate on the data requirements for the two commonly used commercial survey sensors, the magnetometer and the EM-61.

In this study, the data-collection specifications were designed to support classification analyses. The array data, for example, were collected with systems designed to collect research-quality data operated by highly trained scientists. Data of sufficient quality could be collected using other contractor-built arrays operated by field personnel, if the data specifications and quality checks were adequate, as demonstrated by the EM61-MK2 CART results.

6.1 DENSITY

For model-based inversion of geophysical data to be successful, a sufficient number of data points must be collected to accurately represent the magnetic field or EM response of the target object. In the case of magnetometer data, which is fit to a simpler model with fewer parameters, the data density required is not as high as it is for EM data. The EM signal is fit to a more complicated model with more parameters, and more data points are required to obtain a fit that converges and provides a physically reasonable answer. Figure 6-1 illustrates this effect with EM61-MK2 data.

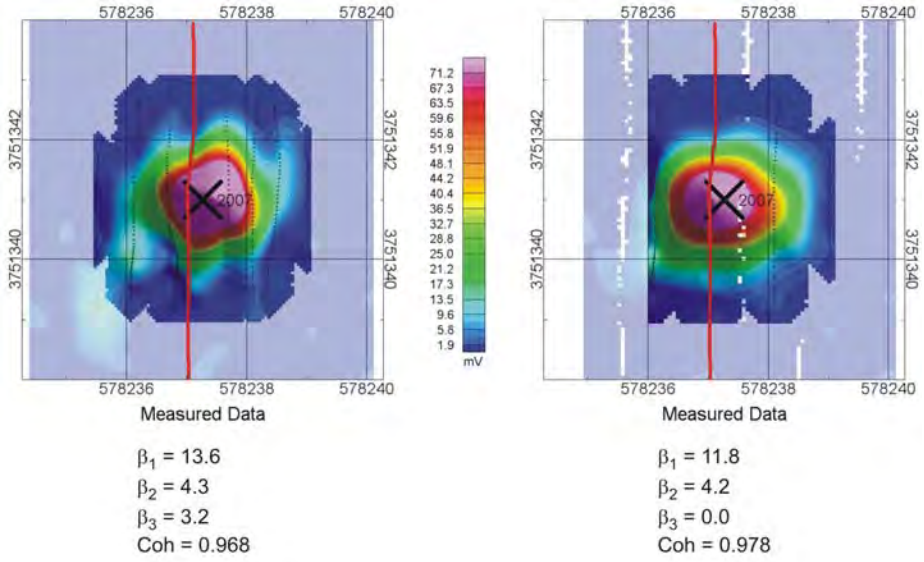

Figure 6-1. Results of fitting EM61-MK2 data that is adequately sampled (left) and that is undersampled (right). Axes' units are in meters. Although the fit quality (coherence) appears better in the undersampled case, the derived parameters are not sensible. The target is a mortar, which is a cylindrically symmetric object. In the analysis on the left, β_2 and β_3 are nearly equal, as they should be, whereas in the analysis on the right, they differ considerably.

Geophysical data are commonly acquired by running a sensor in closely spaced lines, similar to the pattern of a lawnmower cutting grass. Sensors can be ganged in an array to collect more lines of data

for each physical pass over the site. In either case, data density is determined by the spacing between adjacent lines, termed lane spacing, and the combination of down-track speed and sampling rate (how frequently a data point is recorded), termed along-track. The cross-track lane spacing almost always exceeds the spacing between points in the down-track direction.

The MAG data were collected with an array system. Eight magnetometers are spaced 25 cm apart to cover a 2-m swath width on each pass. Adjacent passes are overlapped slightly so that no line spacing exceeds 25 cm. The sampling rate is 50 Hz, which, combined with a typical down-track speed of 3 m per second, results in a measurement in the down-track direction approximately every 0.06 m.

The EM ARRAY data were taken with an array of three overlapping 1 m^2 EM61-MK2 MK2 sensors that cover a 2-m swath. The three EM sensors are configured so that the data line spacing is 0.5 m. The sampling rate is 10 Hz, which, combined with a typical down-track speed of 1.5 m per second, results in a measurement in the down-track direction approximately every 0.15 m. Because the EM fit requires higher data density and benefits from a diversity of measurements from different orientations relative to the target, the field was surveyed twice in orthogonal directions.

The EM CART data were collected with a typical industry-standard 0.5 × 1 m EM61-MK2 coil. The ongoing production survey work at Camp Sibert to support detection specifies 1-m lane spacing. For the pilot project, this was reduced to 0.5 m in the cross-track direction. The measurements in the down track direction were taken approximately every 0.06 m.

6.2 GEOLOCATION PRECISION

Another key element to obtaining a good fit of data to a model is that the locations at which the data points are recorded are well measured. It is important that all the points to be incorporated in the analysis of each target be well located relative to one another, that is, that they have high precision. The absolute accuracy of the point locations in a global coordinate system is less important. For MAG fits, relative errors on the order of about 10 cm or less are tolerable. For EM fits, tolerable errors are smaller, ideally 1-2 cm, but necessarily 5 cm or less. Figure 6-2 illustrates the degrading effect that increasing location error imposes on the parameter value estimates from EM modeling. Needed precision can be achieved under good conditions with DGPS.

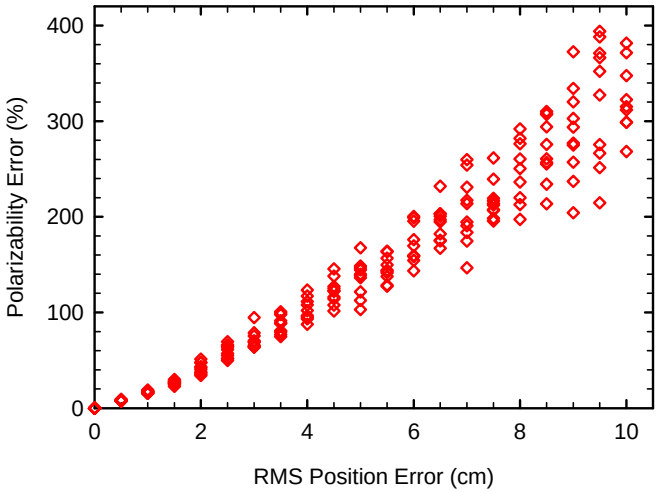

Figure 6-2. EM polarizability errors as a function of position error.

In addition to the absolute location of a GPS antenna (or other positioning measurement device), it is advantageous to measure the orientation of the platform. This allows for correction of any sway in the GPS antenna. If the antenna is mounted on a long pole, as shown in Figure 6-3, this effect can introduce errors of tens of centimeters in the apparent sensor position, which would be detrimental to the model fitting.

Figure 6-3. Error introduced by uncorrected tilt of GPS antenna.

For EM61-MK2 and GEM-3 array systems, geolocation was done in the same way. The platform contains three GPS receivers and an IMU. The data from the three GPS sensors give an accurate location in three dimensions, which can be combined to provide some information about the platform orientation. However, the GPS sample rate is too slow to capture all platform motions, so an IMU with a much faster sample rate is used to supplement the orientation information. (Ref. 11) The magnetometer array position is determined by a single GPS antenna.

The cart data were positioned with a centimeter-level GPS system as it is typically used by the contractor. A single antenna was mounted on a pole over the center of the coil. No orientation information was captured.

6.3 SIGNAL-TO-NOISE RATIO

There are other data-quality considerations related to the SNR that are important for obtaining data that are useful for model fitting. Only data with sufficient SNR can be successfully modeled. Estimated minimum SNR for a successful model fit varies from about 4 to 10. Both the system noise and the SNR for a calibration target should be specified and regularly checked. The actual values that will be appropriate will depend on the site conditions and the munitions of interest. Potential problems and their effects are briefly summarized below, along with considerations for mitigating problems.

Signal: To extract physically meaningful parameters, it is essential to ensure that the sensor measurements are accurate and repeatable. If the signal measured for the same target varies from hour to hour or day to day, it is not possible to use the extracted parameters to make a classification decision. In this program, the response of the sensors to calibration targets was measured for consistency at the start and finish of each data-collection day to provide a quantitative measure that

the sensor was working properly. To mitigate the tendency to be more careful collecting data for calibration than during typical field operations, the signal strength for blind seed targets can also be checked for reasonability. The measured values should fall within the predictions discussed in Chapter 4.

Noise: There are numerous sources of noise that would result in data that is not analyzable. These include noise from the platform, vehicle, operator, and external magnetic sources, as well as from poor sensor condition. Noisy data can either result in a fit that will not converge or in parameters that are contaminated with extraneous contributions to the signal that is being modeled. A site-specific noise specification should be developed and checked regularly. A target-free area can be used to determine normal background noise conditions. This may, of course, vary across the site. Each day's data should be examined to verify that the measured noise is as expected or to document differences across the site so that they may be incorporated into the analysis and decision-making processes.

6.4 DATA ACQUISITION OBSERVATIONS AT SIBERT

Despite strict data specifications and careful field work, problems were encountered with the data collected during this demonstration. Here we discuss those issues that affected data quality and make some observations about platform selection in general. Fortunately, none of these problems was serious enough to compromise the classification study as a whole, but effects were observed in reduced performance on some data sets and in failures to correctly classify specific targets that were affected.

Furrows: Part of the demonstration site had been recently plowed, leaving furrows in the east-west direction. As the data-collection platforms moved over these furrows, a regular pattern of motion-induced noise was introduced. This was particularly a problem for the EM61-MK2 ARRAY. An example of the data is shown in Figure 6-4. The data collected in north-south passes shows regular motion-induced noise, but that collected in the east-west direction does not.

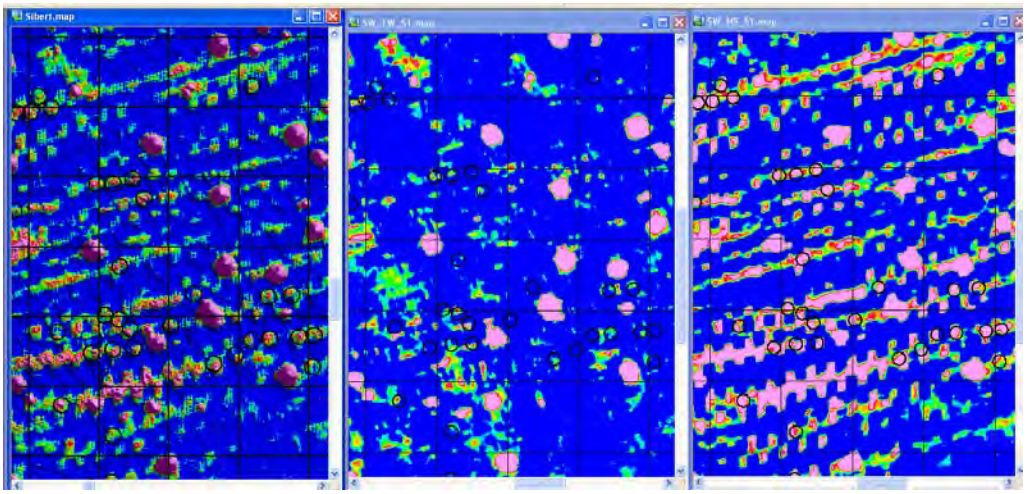

Figure 6-4. Platform motion noise in the EM ARRAY data. The left panel shows the combined north-south and east-west surveys, the center panel shows the east-west survey alone, and the right panel the north-south survey alone. Furrows from recent plowing introduced regular noise features in one direction.

GPS drop outs: GPS systems occasionally experience drop outs. During the quality-control (QC) process, the GPS data are examined for fix quality and the data are plotted for visual inspection of x and y. However, there was one small section of data collected where the z location value was not accurate. Since the model fitting is a three-dimensional operation, the incorrect z resulted in nonsensical fits for a handful of targets using this data set. This was discovered during the failure analysis conducted after the classification results had been scored.

Lane following: Gaps in data coverage can result in too few data points to accurately characterize a target. Such gaps are often the result of small errors in following the planned lane pattern. Long lanes can be difficult to follow exactly from visual cues, such as cones marking the ends, alone. This was a problem in a few places with the cart data. Real-time GPS-driven navigation systems were used to guide the array survey systems. These navigation systems import the planned lanes and report the position and any error to the operator so that track corrections can be made in real time and planned lines closely followed.

Array versus single sensor: Where it is applicable, there are a number of advantages to using an array. In addition to the increased production rate from collecting a wider swath of data in a single pass, the array setup provides data that are better suited to feature extraction. Primarily, this is because within a swath width the data are well located in a relative sense. For the magnetometer array, for example, the eight magnetometers are rigidly spaced and all the sensors on the platform move together. Even if the platform is slightly mislocated, the locations of the sensor measurements relative to one another are well known. In addition, the array decreases coverage mistakes from a narrow-swath sensor following lines of approximately the same spacing as its width. Finally, the greater weight and power that can be borne by a towed platform allow for more complex geolocation systems and navigation aids than might be practical in a man-portable configuration.

For EM systems, the collection of array data needs to be carefully planned. The ability to extract parameters from EM data depends on sampling the signal from multiple orientations of the sensor relative to the target. When EM sensors are arrayed and the transmitters are fired simultaneously, this spatial diversity in the illumination of the target is reduced. This was compensated for by collecting data in two passes run in orthogonal north-south and east-west directions. Since the transmitter is moved between each data point in a single-coil system, this is not an issue for the cart system.

7 COST CONSIDERATIONS

It is premature to attempt a quantitative cost assessment based on the Sibert study alone. The demonstration nature of the project required extensive planning that resulted in a very conservative data-collection plan and the collection of redundant data. Most data collection was performed by researchers and other senior personnel. In addition, the site was relatively small, so true production costs were not realized even for the more routine data-collection platforms. For the most part, the processing was performed by the developers of the software in their first real-world demonstration, so there was a significant learning curve involved. In short, the costs of the demonstration do not reflect what could be accomplished using a survey crew and experienced geophysicists from a competent contractor to collect and process the required data. This is achievable and production costs will likely be substantially lower than those in the demonstration.

7.1 COST MODEL

The cost associated with characterizing and remediating a munitions response site is driven by many factors. In current site-characterization operations that use digital geophysics, all items that are detected by a sensor above a target selection threshold are dug. Many of the excavated items above these thresholds are not munitions and include non-munitions-related man-made items, natural geology, or nonhazardous munitions-related scrap.

Currently, trained explosive safety personnel must be utilized for each dig at a munitions response site. On some sites, explosive safety exclusion zones require the use of barricades or evacuations or prevent the simultaneous deployment of multiple dig crews. In the case of chemical munitions sites, elaborate safety and monitoring equipment must be used. Until classification has been demonstrated successfully on a substantial number of sites to build confidence, it is possible that leaving items completely unexcavated may not be accepted by stakeholders. In this case, if all detected items are to be dug, site managers could utilize classification techniques and employ less expensive procedures, under the supervision of a UXO supervisor, to dig all the items that were identified as not hazardous.

Two key site-management decisions drive classification costs: the threshold for anomaly selection and the use of classification to decide how to treat each anomaly. Objective-driven detection, coupled with classification approaches to identify with high confidence items that are not munitions, has the potential to dramatically reduce the overall costs of a remediation project. We have developed a notional cost model to show how the tradeoffs will occur.

The cost model considers three main elements:

- Data collection costs, since data required for classification may cost more to collect,
- Additional processing costs, and
- Savings from digging fewer holes or not having to use full-up safety measures.

The model is intended to apply to survey data acquired using commercial instruments that could be contracted today. Cued data and research instruments may have substantially different cost structures and are not considered explicitly.

7.1.1 Cost Savings From Anomaly Selection Threshold

Traditional methods to select the target-detection threshold have not been rigorously tied to the munitions of interest. One common method has been to tie the threshold to the noise floor seen in the GPO, often setting it at a multiple of 1.5 times the measured noise. This has the effect of introducing a large number of anomalies arising from noise and small clutter into the target lists, despite the fact that their signal strengths are not consistent with the targets of interest.

This study used as the threshold the signal strength calculated for the 4.2-inch mortar at the deepest depth and the least favorable orientation, divided by two as a safety margin. Figure 7-1 shows for one of the EM61-MK2 data sets the number of anomalies that would appear on the detection list for a threshold set:

- At the noise floor (3195),
- At 1.5 times the noise floor (2541), and
- Using the predicted munitions signal strength (997).

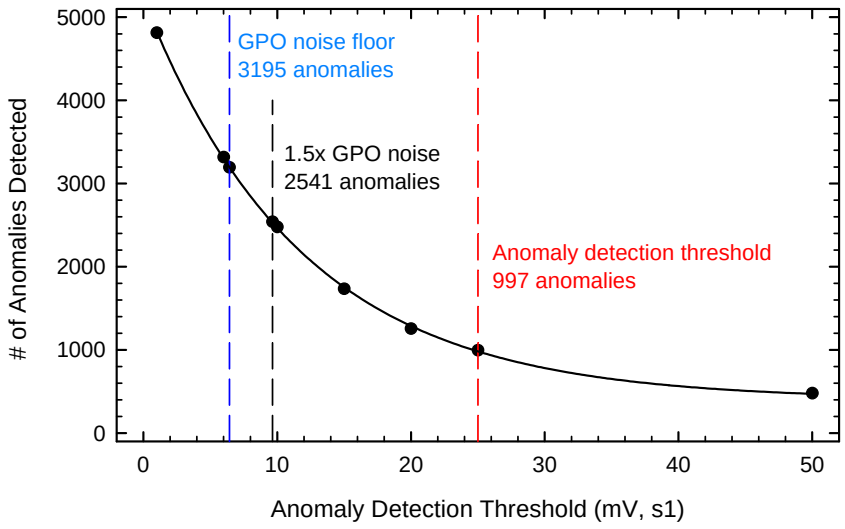

Figure 7-1. Anomaly detection threshold used in this study compared to traditional methods. Note that ~1500 additional anomalies are selected at a threshold 1.5 times the noise compared to the method used in the pilot study.

This method detected all the seeded munitions and significantly reduced the number of items requiring further action. In the common method of setting the threshold to 1.5 times the noise, all these additional items would require excavation. In this method, it is recognized that these signal cannot be the result of the target of interest and they are neither excavated nor further analyzed in the classification processing. This is significant because the majority of these items would have likely have had insufficient signal to be analyzed and therefore been placed on the "can't analyze" list and excavated. In addition, the cost savings from this step alone would be 1544 times the cost to dig each hole.

7.1.2 Cost Savings from Classification

In addition to selecting an appropriate detection threshold, determining which of the detected anomalies can be treated as high-confidence clutter and which must be treated as potential munitions is an important cost driver. The classification approaches applied to this site were used to produce hypothetical examples of how the cost trade-offs would work if classification were implemented at an operational site. No attempt is made to make a quantitative cost estimate, but general trends and break points are noted.

In this study, all detected anomalies were sorted in a list from highest confidence that the item is clutter to highest confidence that the item is a munition. Classification technology demonstrators had to identify the breakpoint on the dig list beyond which all items could not be determined to be clutter with high confidence and must therefore be treated as potential munitions. All items that could not be analyzed with the classification approaches were automatically placed on the list that must be treated as munitions.

The lower panel in Figure 7-2 shows how notional costs accumulate through the process of data collection and processing, digging the munitions, and digging the clutter. The examples presented in Figure 7-2 compare costs based on three scenarios:

- Status Quo Detection Only: This model specifies a lower density data collection for detection only and all anomalies are excavated using intrusive recovery procedures that require trained UXO qualified personnel and safety equipment.
- Classification 1: This model specifies higher density and higher quality data collection followed by classification processing, and all high-confidence clutter items are left unexcavated.
- Classification 2: This model also specifies higher density and higher quality data collection followed by classification processing, but a less expensive alternative to the current operational methods of intrusive recovery is used on the anomalies determined to be clutter with high confidence.

The classification examples are tied to the different regions of the ROC curve shown in the upper panel. For a detailed explanation of the ROC curves, see Section 2.

There are several important points to note in interpreting this plot:

- The cumulative cost curves start out on the *y*-axis at different points. This reflects that the initial costs of higher density data collection and processing for classification are higher than the standard methods. The costs of digging the munitions, which must be borne in all cases, are included here.
- The "detection only" curve has a constant slope and ends at the total number of anomalies. All detected anomalies are dug using the same procedures at the same costs.
- For both classification examples, all the items determined to be high-confidence munitions or "can't decide" must be dug as though they are munitions. Thus, the two classification examples rise at a slope equal to the detection slope until the threshold is reached on the ROC curve where clutter is identified with high confidence (i.e., the yellow-green transition).

- In the region where there is high confidence that the remaining anomalies are clutter (green portion of the ROC curve) and it is decided not to dig these anomalies at all, no additional costs are incurred.
- In the region where there is high confidence that the remaining anomalies are clutter and it is decided to dig these anomalies, but using alternative dig procedures, additional costs are incurred, but the cost of each of these digs is lower so the slope is more gradual.
- The break point in cost saving will be determined by the true dollars associated with the data collection, processing, and digging, which will be site specific.

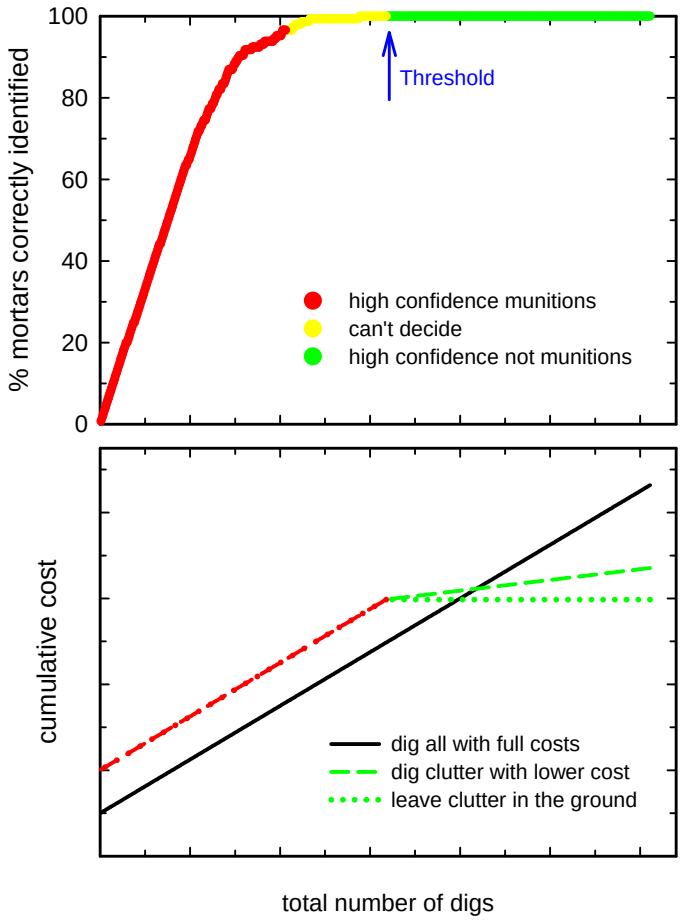

Figure 7-2. Notional cost model illustrating the potential savings using the classification methods outlined in this report.

7.2 COST DRIVERS

Many factors will affect the actual real-world values for our three main cost elements. The breakpoint in cost savings will be determined by the true dollars associated with these elements. Here we discuss the important factors that will influence these values.

Data collection: The costs for a production application of these technologies are dependent upon site conditions such as topography, vegetation, geologic background, and known munition types.

These factors will determine the types of sensors that may be used, the platforms that will be appropriate, and the care with which they must be operated.

In the study at Camp Sibert, the array systems surveyed 100% of the study area, which is amenable to low-cost, high-quality data collection. In the case of the magnetometer array, the magnetometers are fixed at a spacing of 0.25 m, which is adequate for classification. Presumably, the costs would be the same for a detection or a classification survey. The EM61-MK2 array, with 0.5-m lane spacing, surveyed the study area in two perpendicular directions. If orthogonal surveys such as this are needed, data-collection costs will be commensurately higher. The EM61-MK2 cart system surveyed the site with 0.5 m lane spacing for classification, where the specification for the production work on the rest of the site to support detection is 1 m.

Processing: Processing costs may be affected by the presence of complex geology, which can make filtering and parameter estimation more complicated. On the Sibert site, the introduction of motion noise in the EM61-MK2 data stream was a factor in the complexity of the data analysis.
The munitions of interest will likely have a great effect on complexity and costs of processing. Here, size was a good indicator, but that will not be the case everywhere. The number of nonmunitions that can be removed with high confidence at another site may be lower. In addition, the job of the processor in determining the important features and training the classifier may be harder. Finally, in operational classification applications, excavation costs will be driven by how conservatively the stop-dig point is placed on the ranked dig list.

Cost of digging a hole: The costs associated with excavating anomalies vary widely. At Sibert, where there are homes adjacent to the site, safety devices known as open front barricades are required for many digs, and this drives the cost to an average of $192 per hole using the prescribed procedures. At the adjacent chemical site, with same types of 4.2-inch mortars, the safety procedures are far more elaborate and the costs per hole are substantially higher. Many sites report much lower costs per hole. When minimal engineering controls are used, costs as low as $45–$90 per hole have been reported. This will be especially true if the costs are dominated by digging small, near-surface junk. These digs would be eliminated at the detection stage based on the threshold used in the Sibert study model. Cost estimates for such sites would need to account for not only the reduced number of digs, but also the differences in the types of targets that would actually be dug and their average cost per target versus current practices. In general, large, deep targets cost more to dig.

8 PROGRAM CONCLUSIONS

8.1 OVERALL

The pilot program demonstrated successful classification on this simple site. With carefully collected survey data from either magnetometers or EM sensors and transitioning physics-based analysis techniques, well over half the detected clutter items were routinely eliminated with high confidence. Some of the demonstrated processing approaches on these data sets were able to eliminate up to 75% of the clutter. When advanced emerging EM sensors were used, nearly perfect results were achieved.

With regard to detection, the pilot program demonstrated an approach that ties the target-selection criteria to the expected signal strength of the target of interest rather than the site noise. For the common survey sensors, the target signal is readily estimated and confirmed with measurements in a test pit or GPO, and is site invariant. This approach alone substantially reduced the number of signals that were selected as potential targets of interest. For example, in the EM61-MK2 Array data, establishing the target selection threshold using the predicted signal strength with a factor of 2 safety margin reduced the number of targets selected to less than 1000, compared with more than 2500 using 1.5 times the site noise.

8.2 ABILITY TO CORRECTLY DETERMINE PARAMETERS

A critical factor in this process is deriving physically meaningful parameters. In this study, the munitions were most commonly identified by parameters related to their size, which was large in relation to the predominant clutter on the site. Other parameters that were important were decay constants related to the wall thickness and material properties of the objects.

Meaningful and consistent parameters were derived not only for the munitions items, but for common classes of clutter items as well. An example is seen in Figure 5-3. Here, the munitions form a cluster and the common clutter items, including base plates, nose cones, and half rounds, do so as well. This adds confidence that the underlying concept of deriving physically meaningful parameters to identify classes of items with high confidence is valid.

8.3 UNDERSTANDING FAILURES

Failures in the attempt to use physics-based analysis to make a decision about whether a signal corresponds to a target of interest can come in two ways. We have termed these *can't analyze* and *can't decide*.

Can't Analyze: Equally important to determining meaningful parameters is the ability to determine when derived parameters are not meaningful and should not be used for classification decisions. Low SNR, poor data quality from any cause, or an inadequate physical model can cause a poor fit. This is evident in a fit quality measure. Most demonstrators were able to analyze 70%–80% of the anomalies. All items for which good fits cannot be obtained need to be treated as unknowns and dealt with as though they were in fact dangerous munitions.

Many of the failed fits corresponded to soil, hot rock, or no contact. It is not expected that a model of a compact metallic object will necessarily fit well to such signals. For example, one analysis turned

out 196 *can't analyze* targets. Of these, 98 were not metallic objects. Although largely benign, parts of the Camp Sibert site exhibited moderate geological interference. At another location, the number of *can't analyze* targets attributable to geology may be higher or lower.

The two main causes of *can't analyze* targets are that the data do not adequately capture the target signal or that the model is physically incomplete. Unless better data can be obtained or a different model that better describes the physics is employed, there is no possibility of improving performance on targets for which meaningful parameters cannot be estimated.

Can't Decide: The other condition where the analysis model demonstrated that it would be unable to make a classification decision is when meaningful parameters are obtained (i.e., the fit converges), but the parameters do not fall neatly into well defined categories that will permit classification. That is, the munitions and the clutter "look" similar to the sensor.

In this case, it is possible that additional training data can improve the ability to make a determination. As items are excavated, their identities can be incorporated into the classifier to improve the statistical characterization of the objects on the site and their properties. If this additional training data sufficiently improves the separability of different classes of items, it is possible that the *can't decide* items can ultimately be classified. In general, the demonstrators also did a good job of setting their thresholds to correctly account for the items where they could not make a high confidence classification decision.

8.4 SITE–SPECIFIC FACTORS AFFECTING PERFORMANCE

The conclusions about classification performance on this site must be interpreted in light of the site conditions. It is not expected that the performance observed here will be transferable to other sites with much more difficult conditions. Additional demonstrations are planned to document performance under more challenging conditions.

Several factors at Camp Sibert favored success. These include:

- Single munitions type: On a more complex site, with a mix of munitions types, classification will be more complicated. The combined parameter spaces that define multiple munitions types will likely overlap with more of the clutter items.
- Large target of interest: Large targets provide several advantages. First, the large targets tend to have strong signals, which yield ideal data for doing good quality parameter estimation. Second, a large item of interest among smaller clutter items makes size an important parameter, and size is well determined by the parameter-estimation methods used. Third, the classification problem is simplified if there is good separation in a single parameter.
- Isolated munitions: It is difficult, if not impossible, to obtain meaningful parameter estimation results on items with overlapping geophysical signatures. That is currently a subject of research. The models used in the pilot program assume that items can be fit to a dipole associated with a single object. If adjacent signals contribute to the observed signal, the parameters that are derived will be contaminated. We deliberately avoided areas where densities were high and many signals would overlap. In addition, we eliminated from the study any isolated groups of items, which cannot currently be treated and must all be regarded as not analyzable.

- Site conducive to taking "good" data: This site was relatively flat, had good sky view for high-quality GPS, exhibited little geologic interference, and was generally conducive to collecting data of a quality that supported parameter estimation with physics-based models.

9 CLASSIFICATION IMPLEMENTATION

Realizing the potential advantages of classification requires formulating a model for its application that will be accepted by all stakeholders. The study described in this report relied on a retrospective analysis of the demonstrators' performance. All anomalies on the various dig lists were excavated and compared to the rankings from each demonstrator to determine performance. This is possible only in a demonstration project, with perfect knowledge after the fact, but is not a good model for an actual use of these techniques, as avoiding some digs is the major benefit of successful classification. In this section we consider how classification might be implemented in the absence of complete and perfect knowledge.

9.1 PRACTICAL MODEL FOR THE CLASSIFICATION PROCESS

Classification on a production site would need to proceed in a prospective rather than retrospective model. Truth information that would allow one to determine whether each individual item was correctly classified as munitions or clutter will not be available if the clutter is not dug up. At the conclusion of the initial analysis and training period on a site, the site team will have the results of the analysis of the training data, a prioritized dig sheet based on that training, and a recommended threshold based on a combination of accumulated experience and site-specific factors. The challenge for the site team will be how to proceed with the dig program, and specifically, how to decide when to stop digging (or when to relax the safety requirements as the targets being dug are no longer hazardous).

One obvious possibility is to accept the analyst's threshold and stop digging at the threshold. For a number of the analyses discussed in this report, that would have been a successful strategy, although not all demonstrators drew their threshold to capture all munitions and this was only known after the fact. If the dig results make sense, that is, if there are lots of munitions dug when the items with high probability of munitions are dug, and none appear when the threshold is approached, this will increase the stakeholders' confidence in the threshold chosen.

There are a number of actions that could be taken to increase confidence in the validity of the suggested threshold. One, or more, of these validation methods will likely be used on all sites.

- ***Seeding the site*** can increase confidence in the classification process if all the seeded items are found and correctly classified. This does not, however, guard against the presence of unexpected classes of hazardous items.
- ***Complete remediation of a small number of grids*** to test the performance of the classifier/threshold combination is another possibility. This would presumably be an extension of how the training data would have been acquired; all anomalies are dug in a representative sample of grids. The remainder of the site can be dug using the threshold validated in this "blind test." As long as the grids chosen for training and testing are representative of the entire site, this validation should give stakeholders confidence in the classification process.
- ***Investigating a sample of items on the do-not-dig list*** can also be used to validate the classification results. It does not make sense to randomly select items classified as nonhazardous for digging as the likelihood of finding one of the few missed munitions among the overwhelming number of clutter items is small. It might well make sense to sample items based on their location in parameter space. In Figure 9-1, a plot is made with

the estimated value of feature 1 on the *x*-axis and the estimated value of feature 2 on the *y*-axis. Each anomaly analyzed is represented by a point on this plot corresponding to the feature values estimated for that anomaly. The two green clusters were expected classes of munitions. The red cluster represents a group of like items that was not expected. It would be prudent to sample the anomalies that make up the red cluster. If, in fact, they were correctly classified as nonhazardous, this will increase the site team's confidence in the process. If, on the other hand, they turn out to be munitions not represented in the training set, then the conceptual site model must be revised and a new munitions type added to the objectives.

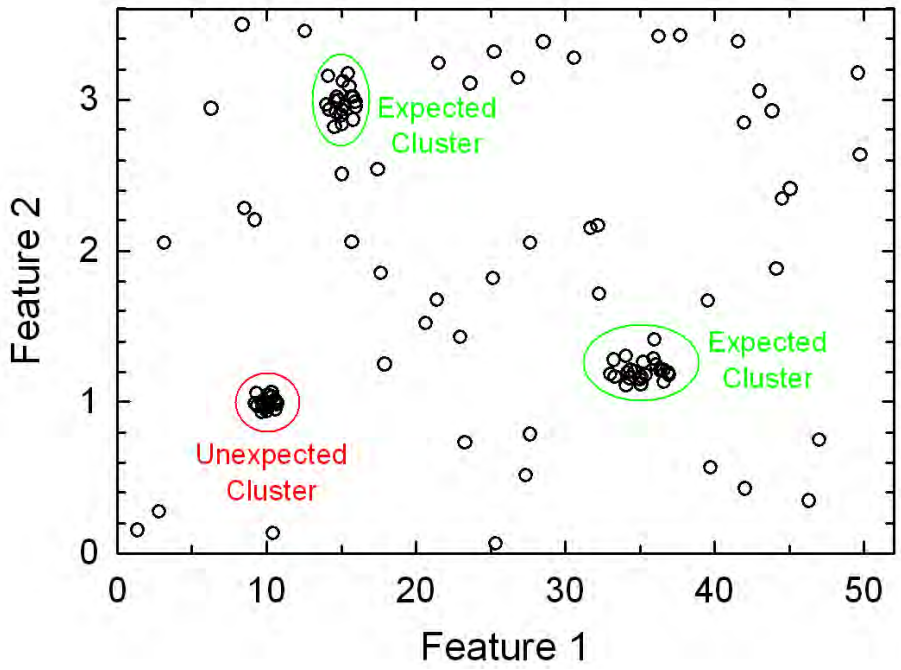

Figure 9-1. Cartoon of a two-feature classification in which two expected clusters of anomaly features are observed, along with one unexpected cluster.

In addition, it must be recognized that the classification process is not static. Each dig provides additional ground truth that can be used to augment the initial training data. It is likely that the classifier will be retrained on a regular basis (nightly if there are several dig teams active or, perhaps, weekly). Over time, this additional ground-truth information will sharpen up the parameter distributions for each item of interest and likely lead to adjusted thresholds. Unfortunately, this additional training data will be preferentially obtained from anomalies classified as resulting from munitions items, which are the best defined initially and therefore the least useful as additional training items. This additional training can, of course, only help with classification of items for which reliable features could be extracted; it can do nothing for anomalies that could not be successfully analyzed.

9.2 Applications of Classification to the Munitions Response Process

The classification process as discussed above has obvious applications in two stages of the munitions response process. One of the goals of the investigation stage is to ensure that the historical records regarding munitions use at the site are correct and complete and to gain some

information about the distribution of contamination across the site. At present, this is accomplished by prioritizing the anomaly list by size or amplitude combined with selective digging. If quality data are collected, even a two-feature classification as illustrated in Figure 9-1 could be used to guide the dig program more efficiently.

In this example, a number of anomalies are scattered throughout the plot and three clusters of anomalies are observed. Two of these clusters, circled in green in the figure, correspond to feature values expected from the two items of interest known from the historical documents. The third cluster, circled in red, was not expected and could correspond to a previously undocumented item of interest or a class of nonhazardous item such as fins. Digging a representative sample of this third cluster will easily distinguish these two possibilities.

The second application of classification techniques is, of course, during a remedial action. As discussed in the earlier chapter, successful classification can lead to more appropriate use of expensive safety measures or, in some cases, the decision to leave items in the ground.

9.3 FACTORS AFFECTING ACCEPTANCE OF CLASSIFICATION

9.3.1 Transparency of the Classification Process

The model for classification outlined in this report represents a transparent process involving explicit, documented classification that should aid acceptance by stakeholders. It should be realized that current survey practices all involve classification, if only implicitly. A mag & flag or EM & flag survey involves operator judgment as to the threshold to declare a detection; this process is neither documented nor reproducible. Even a digital geophysical survey typically involves an analyst examining the geophysical data and making some determination of what to include on the anomaly list; this is often not documented but can be reexamined.

The process that has been outlined here is documented at each step, is reproducible, and can be re-done if site objectives change. The pre-processing steps applied to the raw geophysical data are documented, the anomaly selection threshold is based on the target(s) of interest at the site, and the classification algorithms and thresholds for dig list ranking are documented. As stakeholders become familiar with this process, their confidence in the results and willingness to sanction its use at their sites should increase.

9.3.2 Quality Assurance/Quality Control Approach

Implementation of a classification protocol as demonstrated in this study will require significant changes to current QA/QC procedures. Current quality concepts will need to be adapted to account for a different expected end state. Many things that are currently considered a QA/QC failure for a detection survey, such as detection of an anomaly using an analog instrument or observation of a significant piece of ferrous metal, are not failures when considered in the classification process. A different QC procedure will have to be devised for the two parts of the process, the detection phase and the classification phase.

QA/QC on the detection phase of this model will require the use of digital instruments, since detection thresholds are set by expected signal amplitude from the target of interest and low-level anomalies will be present at the end of the procedure. This could be accomplished by surveying

small portions of the site with an instrument comparable to that originally used for the detection phase. The use of seed targets in the site can also provide valuable detection verification.

QA/QC criteria need to be developed for the classification phase of this model. If all anomalies above the threshold are removed and classification is used only in selecting the level of safety procedures required, then the information is available for detailed QC of the classification results. If, on the other hand, anomalies classified as nonhazardous are left in the ground, large metal objects may be legitimately remain in the field. Procedures for QC/QA will likely include:

- Seeding the area with inerts, which can be used not only to verify the ultimate decision for each seed target, but also as a check on the parameter-extraction process for convergence and meaningful parameter estimation, and
- Selective, guided digging of targets classified as nonhazardous as discussed in Section 9.2.

9.3.3 Management of Residual Risk

Implementation of these techniques should not, in itself, change the way that residual risk is managed at munitions sites. Fully implementing the power of classification, that is, leaving items classified as "not munitions" in the ground, introduces a small risk that a munition may have been incorrectly classified and remains on the site. This is no different than the small probability that a munition may not have been detected using current techniques and therefore remains on the site. In both cases, stakeholders must realize that there is no way to reach 100% certainty that a site is munitions free. Residual risks must be appropriately managed in both cases.

10 FREQUENTLY ASKED QUESTIONS ABOUT CLASSIFICATION

What is classification?
Classification is a process that differentiates munitions from nonhazardous buried items by applying mature physics-based analysis methods to sensor data. Nonhazardous items could include munitions-related scrap, geology, and cultural clutter. The analysis methods are used to estimate parameters of buried objects, such as size, aspect ratio, remanent magnetism, and electromagnetic decay rates, that can be useful in distinguishing munitions from other sources. Advanced classification algorithms then use this information to determine whether a signal is likely to arise from a munitions item or another source.

The results from the classification algorithms are used to develop a dig list ranked from highest confidence the items are nonhazardous to highest confidence the items are munitions. A point on the dig list, termed the threshold, is selected to separate items that need not be treated as potential munitions.

What about other anomaly prioritization methods?
The approach demonstrated in the pilot program represents a physics-based, principled, transparent, process consisting of component methods that are well understood and have undergone several years of development and review by DoD and other interested parties. Alternative anomaly-prioritization methods have been attempted. ESTCP does not have a detailed understanding of these methods, and many differ considerably in approach. Only those methods referenced here were demonstrated as part of the pilot program. The same steps could be implemented by other analysts with the same result. The successes seen in the pilot program should not be expected to transfer to alternative and undemonstrated methods.

How do you evaluate whether a proposed classification approach is real? How do you filter bold claims?
Proposed classification schemes should be physics-based, principled, and transparent, with a set process for quantitative decision making. They should have a development history that provides well-documented and predictable results in controlled experiments, such as test stands and test sites, followed by comprehensive, well-documented demonstration of success at a real field site. So-called black-boxes in which the methods and decision-making criteria are not transparent should be viewed with skepticism.

What about munitions constituents?
None of the classification work demonstrated in this program is applicable to munitions constituents.

There is always some concern about items not being detected. What affect does this have on implementing classification?
Detection and classification should be thought of as separate sequential steps. If an item is not detected, there is no opportunity to classify it. The classification step is performed only on those signals selected as detections, and applying classification can neither improve nor hinder the detection step.

Where could classification fit into the regulatory/munitions response process?
There are several applications for the information gained from classification on a site. Site managers can use the ranked dig list to select an appropriate point to stop digging and leave the remainder of the detected items in the ground.

The results of classification could also be applied to a site where all items will be dug. It is expensive to deploy UXO technicians and in some cases expensive shielding equipment and exclusion zones to support digging of all detected items at a site. The ranked dig list produced from classification could be used to manage dig teams by deploying these measures only where they are necessary.

In the Site Inspection or Remedial Investigation phase, where the objective is to determine the nature and extent of the contamination, classification can be used to guide investigative digging. Sampling items with a variety of physical parameters can lead to a more complete understanding of the site.

What type of sites is classification applicable to?
The pilot project has only validated classification technology on a simple site with a single munition type, benign to moderate geology, limited vegetation, and flat terrain. The influence of site characteristics such as multiple munitions types and interfering geology will be tested in a continuing ESTCP effort that will span several years. The objective is to further define types of sites where classification would be appropriate.

How do I perform classification on my site?
Commercially available technologies were tested as part of this program and showed substantial classification performance. The first step is to collect 100% coverage digital geophysical data over the survey area. For example, EM61-MK2 data can be collected with a cart using tighter lane spacing than would be currently used for a detection-only survey (0.5-m spacing in this case), but the process uses normal commercial data collection and operator procedures. A detection list is generated from the sensor data set and an appropriate threshold is chosen. For this program, the threshold was set based on the signal from the munitions target of interest at a depth 11 times the diameter, the *de facto* expectation for detectability with modern geophysical equipment. A 50% safety margin was applied to the modeled signal at the item's least favorable orientation.

The detected anomalies can be analyzed using the commercially available UX-Analyze module of the Geosoft software Oasis montaj. Each anomaly can be analyzed to extract features such as size, depth, aspect ratio, and electromagnetic decay rate. Classification algorithms then use these features to assign a probability that the item is a munition or nonhazardous. This information is used to create a ranked dig list that orders all anomalies from highest confidence an item is nonhazardous to highest confidence an item is a munition. A point on the dig list can be selected by a project team to identify which items must be treated as potential munitions based on the site-remediation objectives. Additional sensor and data analysis technologies were successful in this project.

What technologies are needed for classification?
Digital geophysical data are required for classification. Both magnetometer and electromagnetic induction data can be used for analysis. Each of these sensors have strengths and weaknesses, and the decision of which sensor to use will depend upon site conditions and objectives. Magnetometers can locate relatively deeper ferrous items, but can only detect ferrous materials, and their effectiveness is reduced by magnetic geology. Electromagnetic induction sensors detect ferrous and

nonferrous metallic objects and can be effective in geology that challenges magnetometers. EM has a limited depth of investigation due to faster signal fall-off over distance than a magnetometer. Care should be taken with data collection to ensure precise sensor location and 100% coverage, which directly impact classification performance.

Data-analysis requires feature extraction (i.e., size, depth, aspect ratio) from each of the detected anomalies and classification of these anomalies by assigning a confidence they are likely munitions or nonhazardous items. This process can be conducted in the commercially available Geosoft software package Oasis montaj as part of the UX-Analyze module.

Why doesn't current practice classify munitions and clutter?
Current practice is motivated by detection, which sets the data requirements and the decision process. There is no reason in principle that the decision process could not be modified to add a classification step. Current commercial instruments as they are deployed by the contractor community can collect data that may be used for classification. In most cases, doing so would require revisiting the data requirements.

Can classification approaches identify individual items within a cluster of buried items?
The current sensors and classification algorithms are effective only on individual isolated anomalies and do not allow reliable parameter extraction and classification of overlapping clustered anomalies. High-density target centers would not be an appropriate location to apply classification. Current classification approaches can be applied to areas surrounding high-density target centers where isolated anomalies are present. For the purposes of this study, anomalies were considered "isolated" if they were 2 m from the closest adjacent anomaly. It is not envisioned that classification approaches will be applied to the target center, but research is being conducted to analyze overlapping signatures that would likely allow analysis closer to the target center than existing methods.

How can I confirm the technologies perform as they are designed to? How can I be confident in the results?
The techniques used in this study to confirm technology performance can be utilized in operational classification applications. A geophysical proveout (GPO) was established on a small representative section of the site to verify detection thresholds for all data collection systems. The intent of the GPO was to verify that the targets of interest are detected at the depths of interest under site-specific conditions at the selected threshold. The GPO consisted of seeded mortar rounds and splayed half-rounds. The burial depths were biased shallow to provide high signal-to-noise training data, though a few rounds are buried to a depth of 11 times their diameter to verify detection by the geophysical sensors. In addition to the GPO, mortar rounds were also seeded in the survey area itself. The use of a GPO and seeding are both recommended in an operational classification application.

All items above the detection threshold were dug in this study to verify the performance of the sensors and the analysis methods. In an operational application, the ultimate objective is to dig only the munitions and leave the clutter. Random digging beyond the threshold is unlikely to be useful. The number of intact munitions present on most sites is very small, and randomly finding a misclassified munition, if it is present, among overwhelming clutter is very unlikely. This will likely be effective in confirming cultural clutter and munitions-related scrap assignments rather than finding incorrectly classified munitions.

A small portion of the site could be dug completely to confirm the performance of the technologies and refine the classification algorithms. Alternatively, selected anomalies on the dig list could be sampled. The selection of these anomalies could be based on sampling various estimated parameters from the inversion step to confirm the results are physically reasonable.

What if data collected for an anomaly cannot support reliable parameter extraction to identify its features?
For the purposes of this study, these items were not analyzed and were added to the dig list. Demonstrators were generally successful at determining when reliable model fits were not achieved.

How do I determine where to set the threshold once a dig list is created?
There are several factors that can influence this decision. It is recommended the project team weigh the site objectives. Factors such as future land use should be considered. This decision process could be iterative and the point could be moved if information is gained from any validation digging that occurs.

Is classification applicable to marine sites?
There is no reason in principle that classification could not be performed on magnetometer or EM data collected at a marine site. The parameter-extraction process for magnetic data would be identical, and the physics that must be accommodated to account for differences in the EM signal is well understood and could be readily implemented. However, no system that can take data to support classification has yet been demonstrated.

Can you do classification with helicopter data?
Classification from a helicopter platform is expected to be very limited. There are existing magnetometer systems that can take high-density, well-located data at low altitude from a helicopter platform. It is possible to analyze these data to obtain estimates of the target size and depth. The success will depend on the size of the targets of interest and the altitude that can be maintained. For EM systems, where the signal falloff is faster than for magnetometers, no system that collects data appropriate to classification has been demonstrated to date. This is expected to be more limited both because the signal falls off faster with the separation of the target and sensor and because the data requirements are more stringent for parameter extraction from EM data.

Can you use this in the center of an impact area?
Classification to date has been demonstrated only on isolated targets. In areas where the target density is very high and many signals will overlap, such performance has not been demonstrated. This is currently the subject of research, and progress is expected in classification of overlaps of two or three signals and in the analysis of a single strong target shielded by small surface clutter. It is not expected that the current processes for classification will evolve to provide acceptable performance in the centers of impact areas, where individual signals are not separable.

How much does classification cost?
After only a single demonstration, which was a learning exercise for all involved, sufficient data do not exist to make a quantitative cost estimate. As with all surveys and data analysis, the cost of classification will depend on the size of the site and its conditions, as well as the objective. Generally, data-collection costs will be higher and processing will be more involved.

How long does it take to collect data and conduct analysis?
Data collection can range from 1 to 20 acres/day depending upon if the system is man-portable or towed. If the data density required for classification is twice that required for detection only, the field deployment can be expected to approximately double. Data-analysis times will vary depending upon the number of anomalies detected and the presence of complicating factors such as geology. For the roughly 1000 anomalies analyzed in this demonstration, the data analysis required less than 1 person-week for the magnetometer data and less than 2 weeks for the EM. As the classification procedures become better defined, this is likely to decrease.

How specialized is this? What contractor qualifications are required?
Some of the algorithms used in this demonstration are available in a beta version of a package called UXAnalyze, which runs in Oasis montaj. The contractor that collected the cart data in this program downloaded this package and used it to perform successful classification. No special qualifications are needed to run this program beyond those usually held by the geophysicist or data analyst typically involved in a project. It is necessary that the individual understand and be able to evaluate data quality, as well as assess whether reasonable answers are obtained.

What emerging classification capabilities should I watch?
Vastly improved classification performance was demonstrated in this study using a multi-axis data collection system called the Berkeley UXO Discriminator (BUD). This sensor is one of several new generation EM sensors that gather substantially more information in the signals because the systems make measurements at a variety of angles over buried objects. While these sensor systems and associated data-analysis approaches are still in development, they show promise in significantly improving classification performance.

REFERENCES

1. *Defense Environmental Programs Annual Report to Congress*, 2007. Retrieved April 16, 2009 from the website https://www.denix.osd.mil/portal/page/portal/denix/environment/ARC/FY2007.

2. Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, *Report of the Defense Science Board Task Force on Unexploded Ordnance*, Washington, DC 2003.

3. SAIC, Inc, *Interim Report, SAIC Analysis of Survey Data Acquired at Camp Sibert, ESTCP Project MM-200210,* 2008.

4. Sky Research, *Demonstration Report, Data Modeling, Feature Extraction, and Classification of Magnetic and EMI Data, ESTCP Discrimination Study, Camp Sibert, AL, Project 200504: Practical Discrimination Strategies for Application to Live Sites*, 2008.

5. Signal Innovations Group, *Final Report, SIG Analysis of Sibert Data*, *ESTCP Project MM-0501*, 2008.

6. Lawrence Berkeley National Laboratory, *Final Report at Camp Sibert - Gadsden, Alabama, ESTCP Project # MM-0437*, 2008.

7. Cazares, Tuley, and May, *The UXO Discrimination Study at the Former Camp Sibert*, *IDA Document number D-3572*.

8. Defense Environmental Restoration Program, *Archives Search Report Findings, Camp Sibert, Site No. I04AL005700*, 1993.

9. Parsons, Inc., *Final Phase III Conventional MEC Engineering Evaluation/Cost Analysis Report, Former Camp Sibert, Etowah and St. Clair Counties, Alabama*, 2008.

10. Parsons, Inc., *Report of Environmental Security Technology Certification Program UXO Discrimination Study Activities, Former Camp Sibert, AL, Contract W912DY-04-D-0005, Delivery Order 16*, 2007.

11. Harbaugh, Steinhurst and Khadr, *Technology Demonstration Data Report, ESTCP UXO Discrimination Study, MTADS Demonstration at Camp Sibert: Magnetometer/ EM61/GEM-3 Arrays,* 2008.

12. Sky Research Inc., *Demonstration Report for Geonics EM-63 Cued-Interrogation Data Collection, Processing and Archiving at Camp Sibert, Alabama, ESTCP Project MM-0504*, 2008.

13. Foley, et al, *Sensor Orientation Effects on UXO Geophysical Target Discrimination*, *SERDP MM-1310 Final Report*, 2006.

ACRONYMS

BUD	Berkeley UXO Discriminator
DoD	Department of Defense
EM	Electromagnetic Induction
ESTCP	Environmental Security Technology Certification Program
GPO	Geophysical Prove Out
GPS	Global Positioning System
IDA	Institute for Defense Analyses
IMU	Inertial Measurement Unit
LBNL	Lawrence Berkeley National Laboratory
MAG	Magnetometer
MMRP	Military Munitions Response Program
NRL	Naval Research Laboratory
QC	Quality Control
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
UXO	Unexploded Ordnance